الجمهورية العربية السورية

جامعة دمشق

كلية الهندسة الميكانيكية والكهربائية

قسم الهندسة الطبية

تدريب نموذج يد طرف اصطناعي على الإمساك باستخدام التعلم المعزز العميق

رسالة مقدمة لنيل درجة الماجستير في الهندسة الطبية الحيوية

إعداد

م. ثريّا جبران محسنه.

إشراف

أ. د. م. رشا كاظم مسعود

إشراف مشارك

د. م. فاتن حافظ عجيب

2025

تصريح خطي

أنا الموقع أدناه أصرح بعلمي الكامل وقبولي:

1. أن كل ما ينتج عن البحث والأطروحة هو ملكية فكرية ومالية بالتشارك مع جامعة دمشق، وأنني ألتزم بأخذ موافقة الجامعة في حال رغبتي بنشر البحث أو الأطروحة أو جزء منها نصا أو مضمونا خارج إطار الجامعة (من دور نشر أو مكتبات أو مواقع إلكترونية وغيرها من وسائل النشر).

2. أنه لا يوجد جزء من هذه الأطروحة تم اقتباسه من عملي آخر أو أنجز للحصول على شهادة أخرى في جامعة دمشق أو أية جامعة أو معهد تعليمي داخل أو خارج الجمهورية العربية السورية.

الاسم والتوقيع

♥الشكر والتقدير♥

﴿وَآخِرُ دَعْوَاهُمْ أَنِ الْحَمْدُ لِلَّهِ رَبِّ الْعَالَمِينَ﴾

♥ أقف اليوم بامتنان عميق أمام نعمة التوفيق التي منحني إياها الله سبحانه وتعالى والتي لولاها لما تكلل هذا العمل بالإنجاز ولعل أصدق ما أعبر به عن هذا الامتنان دعاء نبي الله سليمان عليه السلام الذي استحضره قلبي ولساني: ﴿رَبِّ أَوْزِعْنِي أَنْ أَشْكُرَ نِعْمَتَكَ الَّتِي أَنْعَمْتَ عَلَيَّ وَعَلَىٰ وَالِدَيَّ وَأَنْ أَعْمَلَ صَالِحًا تَرْضَاهُ وَأَدْخِلْنِي بِرَحْمَتِكَ فِي عِبَادِكَ الصَّالِحِينَ﴾ (سورة النمل: 19). هذا البحث ما هو إلا مسعى لرضا الله ورسوله وطلب العلم عملا بقوله تعالى: ﴿وَقُلِ ٱعْمَلُواْ فَسَيَرَى ٱللَّهُ عَمَلَكُمْ وَرَسُولُهُ وَٱلْمُؤْمِنُونَ﴾ (التوبة: 105)، فليكن هذا العمل شاهدا على سعيي وأدعو الله أن يجعلني من الذين يستمعون القول فيتبعون أحسنه. فالحمد لله أولا وآخرا، ظاهرا وباطنا، على ما أنعم وأعان، والصلاة والسلام على أشرف الأنبياء والمرسلين، وأسأل الله أن يكون هذا العمل خطوة مباركة في طريق العلم.

والشكر موصول إلى من يسر الله وجودي بينهم ووجودهم في دربي فكانوا خير موجهين وناصحين:

♥ إلى مصدر العلم والمعرفة جامعة دمشق العريقة...

♥ إلى البيئية المعطاءة التي احتضنتني واحتضنت بحثي هذا كلية الهندسة الميكانيكية والكهربائية في جامعة دمشق والتي قدمت لي الكثير...

♥ إلى من صقلوا مهاراتي بخبراتهم الثمينة ومنحوني ثقة ومعرفة سيكون لها أثر عميق في مستقبلي المهني والأكاديمي كادر قسم الهندسة الطبية في جامعة دمشق...

♥ إليهم إلى الدكاترة المشرفين على هذا البحث:

♥ إلى من رافقتني بخبرتها وتوجيهاتها القيمة والهادئة وتابعت معي هذا البحث خطوة بخطوة وصقلت رؤيتي البحثية بصبر وإخلاص، مشرفتي د. رشا مسعود والتي كان لدعمها وثقتها أثر عميق في إنجاز هذا البحث.

♥ إلى كلماتها الداعمة وثقتها الدائمة ومساهمتها في إنجاز هذا البحث د. فائق عجيب.

♥ كما لا يفوتني أن أشكر د. رفيدة حمص على اهتمامها في البحث والتي لم تبخل بعلمها وتوجيهاتها الداعمة.

♥ أخيرا، جزيل الشكر إلى أساتذتي الموقرين في لجنة المناقشة د. مصطفى الموالدي و د. سمير كرماش لتفضلهم بقراءة العمل وتقديم ملاحظاتهم القيمة التي سدّت مكامن النقص التي اعترته.

لكم محبتي وأسمى آيات الشكر والتقدير والاحترام..

ثريا جبراي محسنه

♥الشكر والإهداء والتقدير♥

♥ بسم الله الرحمن الرحيم والحمد لله رب العالمين الذي جعل لكل شيء قدر ولكل قدر أجل ولكل أجل كتاب.

♥ ولمن قال في رسالته: "سلُوا الله علماً نافعاً، وتعوَّذوا باللهِ من علمٍ لا ينفع" ذكرتك الروح حباً وعليك القلب صلى.

♥ إلى جيشي الوحيد والأعظم... مصدر قوَّتي وأملي وإلهامي... مبعث الدفء ومنبع السعادة... لمن آمن بي أولاً... وتفانى في سبيل راحتي... من كان دُعائهم سر نجاحي... إلى من بلغ حنانهم مبلغاً لا يضاهى... إلى كيان من حب... إلى أمة في شخصين...

♥ بابا: جُبران محسنه & ماما: مُنى عجال.

♥ إلى شريكتيّ الروح السرمديتين وهبة العمر... مخبأ أسراري الأزلي ومبعث آمالي الأبدي... إلى راحة البال وحلو الحياة... القلوب الطاهرة والنفوس البريئة... نجوم الأُنس في الكدر... كواكب البهجة في الفرح... إلى كل تلك اللحظات والمشاعر التي تقاسمناها...

♥ أخَتَايَ: نُعمى محسنه & بُشرى محسنه.

♥ أنا التي اعتقدت أن خطواتي لا تكفي لهذا الطريق المُتعِب، وأدهشتني عزيمتي..
أهدي ثمرة جهدي هذا إلى نفسي التي تستحق، وأعدها أن أظل مؤمنة بها دائما..
إلى الكثير الذي مررت به، إلى أحلامي النابضة رغم التعب، إلى عثراتي التي صنعتني...

♥ بكل ما تملك روحي من حب... إليّ...

♥ إلى كل تلك الأرواح التي رماها القدر في طريقي فأغنت مسيرتي وأضافت الكثير إلى مشواري الفتي...
إلى محاربي الأمراض في كل مكان، مع أمنياتي بالشفاء والقوة...
إلى كل من يقرأ خلاصة بحثي هذه...

♥ إليكم...

"لو أن الناس كلما استصعبوا أمرا تركوه، ما قام للناس دنيا ولا دين"

ث. ثُريّا جبران محسنه

فهرس المحتويات:

1. الفصل الأول: الإطار العام للبحث 1

1.1. تمهيد 2

1.2. مشكلة البحث 3

1.3. فجوة البحث 4

1.4. أهداف البحث 5

1.5. الدافع وراء البحث 6

1.6. فرضية البحث 6

1.7. أهمية البحث 6

1.8. مساهمات البحث 7

1.9. تحديات البحث 8

1.10. هيكل الرسالة 10

1.11. الخاتمة 10

2. الفصل الثاني: الدراسات السابقة ذات الصلة 11

2.1. تمهيد 12

2.2. دراسات تعتمد التغذية العكسية اللمسية 12

2.3. دراسات تعتمد التغذية العكسية البصرية 15

2.4. دراسات تعتمد أنواع أخرى من التغذية العكسية 18

2.5. التعلم المعزز العميق والتغذية العكسية اللمسية في بيئات أكثر تعقيدا 22

2.6. دراسات تعتمد على التعلم المعزز العميق مع العروض التوضيحية 23

2.7. خاتمة 27

3. الفصل الثالث: اليد البشرية والأطراف الاصطناعية: (دراسة بيولوجية ووظيفية لآليات التحكم والإمساك) 28

3.1. تمهيد 29

3.2. يد الإنسان (دراسة بيولوجية) 29

3.2.1. نظام الإدراك والمعرفة في اليد 29

3.2.1.1. الاستشعار اللمسي في اليد 29

3.2.1.2. الاستشعار البصري وآلية معالجة الدماغ لبياناته دعما للتفاعل اليدوي مع الأشياء 30

3.2.1.3. تعديل الرؤية المرتكز على اليد 31

3.2.1.4. آلية معالجة اللمس في الجهاز العصبي البشري (مسار إشارات اللمس والوظيفة) 32

3.2.1.5. الاندماج البصري اللمسي في الدماغ أثناء التحكم باليد 33

3.2.2. النظام الحركي لليد 33

3.2.2.1. المسار الحسي الحركي 34

3.2.2.2. الهيكل العظمي والعضلات والأوتار 35

3.2.2.3. نمط التحكم في اليد 36

3.2.2.4. ترميز الحركة 37

3.2.2.5. التحكم عبر تنسيق المفاصل المتعددة 37

3.2.2.6. التحكم الهرمي والاستقلال الجزئي 37

3.3. الإمساك وتصنيفه 38

3.3.1. تعريف الإمساك 38

3.3.2. تصنيف الإمساك 38

3.4. البتر والأطراف الاصطناعية العلوية 39

3.4.1. البُتر والأطراف الاصطناعية 39

3.4.2. الأطراف الاصطناعية العلوية وأساليب التحكم والإمساك الخاصة بها 40

3.4.3. أساليب نقل خصائص الإمساك البشري إلى الأيدي الروبوتية والاصطناعية 42

3.4.3.1. النهج التحليلي والمعروف أيضا باسم النهج الهندسي 42

3.4.3.2. المناهج المبنية على البيانات والمعروفة باسم المناهج التجريبية 43

3.5. الخاتمة 43

4. الفصل الرابع: الأسس النظرية للبحث 44

4.1. تمهيد 45

4.2. مفاهيم مهمة 45

4.2.1. الذكاء الاصطناعي (Artificial Inelegant – AI) 45

4.2.2. تعلم الآلة (Machine Learning – ML) 45

4.2.3. الشبكات العصبية الاصطناعية والخلايا العصبية الاصطناعية 46

4.2.4. خوارزمية تدريب الشبكة العصبية الاصطناعية 48

4.3. التعلم المعزز (Reinforcement Learning – RL) 49

4.3.1. نبذة تاريخية عن RL 49

4.3.2. تعريف RL 50

4.3.3. صياغة التعلم المعزز كعملية قرار ماركوف (Markov Decision Process – MDP) 51

4.3.4. العناصر الأساسية لوكيل التعلم المعزز 52
4.3.5. مناهج التعلم المعزز الكلاسيكية..... 54
4.3.6. خوارزميات التعلم المعزز..... 55
4.3.6.1. مفهوم ON–Policy و OFF–Policy..... 56
4.3.6.2. مفهوم أسلوب الفاعل – الناقد Actor –Critic..... 57
4.3.7. مفهوم الاستكشاف والاستغلال Exploitation and Exploration..... 59
4.3.8. أساليب أمثلة السياسة Policy Optimization..... 59
4.3.8.1. خوارزمية تدرج السياسة Policy Gradient..... 59
4.3.8.2. خوارزمية منطقة الثقة لأمثلة السياسة TRPO – Trust Region Policy Optimization . 60
4.3.8.3. دالة الهدف البديل المقطوعة Clipped Surrogate Objective..... 61
4.3.8.4. معامل العقوبة التكيفي Kullback–Leibler – KL..... 63
4.3.9. خوارزمية أمثلة السياسة القريبة (Proximal Policy Optimization – PPO)..... 64
4.3.10. التعلم العميق Deep Learning – DL..... 65
4.3.11. التعلم المعزز العميق Deep Reinforcement Learning – DRL..... 66
4.4. تعلم المناهج Curriculum Learning – CL..... 66
4.5. أهمية المحاكاة في تطبيقات التعلم المعزز..... 66
4.6. مفهوم فجوة الواقع..... 68
4.7. الخاتمة..... 68
5. الفصل الخامس: منهجية البحث..... 69
5.1. تمهيد..... 70
5.2. منهجية البحث..... 70
5.3. تحديد أدوات البحث..... 71
5.3.1. العتاد الصلب 72
5.3.1.1. الحاسوب المستخدم 72
5.3.1.2. اليد التي أجريت عليها التجارب..... 72
5.3.2. العتاد البرمجي..... 73
5.3.2.1. برنامج المحاكاة 73
5.3.2.2. لغة البرمجة..... 75
5.3.2.3. إطار DEEPBOTS..... 75

.5.3.2.4 خوارزمية أمثلة السياسة القريبة Proximal Policy Optimization – PPO .. 77

.5.4 بناء بيئة المحاكاة..... 77

.5.5 حالتي الدراسة (Case Studies)..... 88

.5.5.1 حالة حساسات اللمس..... 88

.5.5.1.1 صياغة المشكلة وفق عملية قرار ماركوف 88

.5.5.1.2 أساليب التقييم..... 90

.5.5.1.3 بناء وتدريب سياسة اتخاذ القرار..... 90

.5.5.1.4 تقييم صلابة السياسة الناتجة وتعميمها (اختبارات المرحلة الثانية)..... 93

.5.5.1.5 تنفيذ تعلم المناهج (اختبارات المرحلة الثالثة)..... 94

.5.5.1.6 اختبار قوة النظام المقترح على إنشاء قبضات مخصصة مستقبلا (المرحلة الرابعة) ... 94

.5.5.2 حالة حساسات القوة..... 94

.5.5.2.1 صياغة المشكلة وفق عملية قرار ماركوف 95

.5.5.2.2 مقاييس الأداء لهذه الحالة..... 96

.5.5.2.3 بناء وتدريب سياسة اتخاذ القرار..... 96

.5.5.2.4 تقييم صلابة السياسة الناتجة 98

.5.5.2.5 تنفيذ تعلم المناهج..... 99

.5.6 الخاتمة..... 100

.6 الفصل السادس: النتائج ومناقشتها..... 101

.6.1 تمهيد..... 102

.6.2 نتائج حالة حساسات اللمس (الحالة الأولى)..... 102

.6.2.1 نتائج اختبارات بناء السياسة والتدريب (المرحلة الأولى)..... 102

.6.2.2 نتائج اختبارات التعميم والصلابة (المرحلة الثانية) 104

.6.2.3 نتائج تجارب اختبار تعلم المناهج (المرحلة الثالثة) 106

.6.2.4 نتائج تجارب اختبار قوة النظام المقترح (المرحلة الرابعة)..... 107

.6.3 نتائج حالة حساسات القوة (الحالة الثانية)..... 108

.6.3.1 نتائج اختبارات مرحلة بناء السياسة وتدريبها (المرحلة الأولى) 108

.6.3.2 نتائج اختبارات تقييم الصلابة (المرحلة الثانية) 114

.6.3.3 نتائج اختبارات تعلم المناهج (المرحلة الثالثة) 118

.6.3.4 نتائج اختبار قوة النظام المقترح على إنشاء قبضات مخصصة (المرحلة الرابعة) 120

6.4. مناقشة النتائج ومقارنتها فيما بينها 123
6.4.1. ملخص نتائج حالة حساسات اللمس (الحالة الأولى) ومناقشتها 123
6.4.2. ملخص نتائج حالة حساسات القوة (الحالة الثانية) ومناقشتها 125
6.4.3. مقارنة النتائج فيما بينها 127
6.5. المقارنة مع الأعمال السابقة 128
6.6. الخاتمة 131
7. الفصل السابع: الاستنتاجات والآفاق المستقبلية 132
7.1. مقدمة 133
7.2. الاستنتاجات النهائية 133
7.3. آفاق مستقبلية وتوصيات 134
7.4. الخاتمة 135
المراجع ..... ا

فهرس الأشكال:

الشكل 2-1 نموذج اليد الأول يمثل اللون الأحمر مواضع المحركات واللون الأزرق المفاصل واللون الرمادي حساسات اللمس (Kepes, 2017). 13
الشكل 2-2 نموذج اليد الثاني المستخدم يمثل اللون الأحمر موضع المحركات والأزرق المفاصل والأخضر حساسات اللمس (Kepes, 2017). 13
الشكل 3-2 الأجسام التي تم التعامل معها أثناء التدريب والاختبار (Kepes, 2017). 13
الشكل 4-2 نتائج التعلم النوعية والكمية ضمن المحاكاة (Ayukawa et al, 2021). 18
الشكل 5-2 المتوسط المتحرك لقيم المكافأة التراكمية عبر الحلقات (Shahid et al, 2020). 20
الشكل 6-2 الرسم البياني للتعلم لمشكلة RL الأصلية مع وبدون IL (Sharif et al, 2020). 24
الشكل 7-2 أداء خوارزميتي NPG وDDPG مع المكافآت المتناثرة والمُشكلة (Rajeswaran et al, 2018). 25
الشكل 8-2 أداء طرق التعلم المعزز مع العروض التوضيحية – DAPG (الخاصة بالدراسة) وDDPGfD (Rajeswaran et al, 2018). 26
الشكل 1-3 المسارات البطنية والظهرية في قشرة الدماغ البصرية والمناطق المتصلة بها (Li et al, 2022). 31
الشكل 2-3 توضيح لعملية نقل الإشارات اللمسية (James, 2021). 32
الشكل 3-3 خريطة (homunculus) (Sinha et al,2019). 34
الشكل 4-3 رسم تخطيطي مبسط للنظام الحسي الحركي. توضح النصوص الخضراء في كل كتلة وظائف المنطقة المقابلة (Li et al, 2022). .. 35
الشكل 5-3 اليد البشرية: (أ) مخطط يوضح عظام اليد البشرية، (ب) نموذج يوضح عدد درجات الحرية لليد البشرية (Redman, 2016). 35
الشكل 6-3 العضلات المحركة لليد: (أ) العضلات المثنية للمعصم والأصابع، (ب) العضلات الباسطة للمعصم والأصابع (Redman, 2016). 36
الشكل 7-3 تصنيف أنواع القبضات تبعا لم ذكره (Kepes, 2017). 39
الشكل 8-3 توضيح لقبضات اليد التي وصفها Schlesinger (Song et al, 2021). 39
الشكل 9-3 مستويات بتر الأطراف العلوية (Cordella et al, 2016). 40
الشكل 10-3 نهج الإمساك المقاد بالبيانات (Bohg et al, 2013). 43
الشكل 1-4 تعلم الآلة مقابل البرمجة الكلاسيكية (Danielsen et al, 2021). 45
الشكل 2-4 العصبون الاصطناعي (أ) نموذج بيرسبيترون (ب) الشكل العام للعصبون الاصطناعي ضمن الشبكة الاصطناعية (Kepes, 2017). 46
الشكل 3-4 شبكة تغذية أمامية مع طبقة مخفية واحدة (Kepes, 2017). 48
الشكل 4-4 النسب المئوية للاهتمام بالتعلم المعزز وخاصة في مجال الروبوتات آخر 10 سنوات (Sekkat et al, 2023). 49
الشكل 5-4 تفاعل الوكيل مع البيئة (Danielsen et al, 2021). 50
الشكل 6-4 خوارزميات التعلم المعزز العميق (Sekkat et al, 2023). 55
الشكل 7-4 الفرق بين التصنيفين (Shahid et al, 2022). 57
الشكل 8-4 موقع أسلوب الفاعل – الناقد بين الأساليب الأخرى. 57
الشكل 9-4 منهجية الفاعل – الناقد (Sutton and Barto, 2018)(Orsula, 2021). 58
الشكل 10-4 خطوة زمنية واحدة للدالة $LCLIP$ كتابع للنسبة r (Schulman et al, 2017). 62
الشكل 1-5 منهجية العمل. 71
الشكل 2-5 نموذج وأبعاد يدHANDY. 73
الشكل 3-5 أكثر لغات البرمجة انتشار في مجال التعلم المعزز (Kepes, 2017). 75
الشكل 4-5 تواصل المتحكم مع المشرف في إطار عمل DeepBots (Kirtas et al, 2020). 76
الشكل 5-5 مقارنة DQN مع PPO من حيث الأنشطة الروبوتية المختلفة، مقابل المكافأة التراكمية (Thompson et al, 2019). 77
الشكل 6-5 توضيح لمفهوم الشجرة والعقد وأطفال العقد في روبوت Aibo. وشجرة المشهد ضمن Webots (Ijspeer et al, 2004). 78

الشكل 5-7 بيئة التدريب..... 79
الشكل 5-8 نموذج يد Handy lite..... 79
الشكل 5-9 صورة ليد Handy Lite ضمن برنامج MeshMixer كملف STL وفرته الشركة..... 80
الشكل 5-10 عرض لشبكة عناصر الجسم ضمن برنامج MeshMixer..... 80
الشكل 5-11 استخدام تعليمات القطع المستوي Plane Cat..... 80
الشكل 5-12 استخدام تعليمة Select –Delete..... 81
الشكل 5-13 السلامية البعيدة للبنصر، يمينا قبل تقليل الشبكة، يسارا بعد تقليل تعقيد الشبكة..... 82
الشكل 5-14 مفهوم تحويل الإطار بين جملتين إحداثيتين Transform Frames (Luo et al, 2023)..... 84
الشكل 5-15 النموذج الذي تم بناءه قبل إضافة المحركات والحساسات..... 84
الشكل 5-16 يوضح الشكل جمع أجزاء الهيكل مرتبطة ببعضها من خلال المفاصل (باللون الأحمر)..... 85
الشكل 5-17 توزع الحساسات على الكف..... 86
الشكل 5-18 بيئة التدريب المبنية ضمن برنامج Webots..... 86
الشكل 5-19 العلاقات الهرمية بين جميع عقد النموذج..... 87
الشكل 5-20 بنية شبكتي الفاعل (أ) والناقد (ب)..... 92
الشكل 5-21 بنية شبكتي الفاعل (أ) – الناقد (ب)..... 97
الشكل 5-22 بيضة الدجاج المصممة في برنامج solidworks واستيرادها لبرنامج Webots..... 99
الشكل 6-1 مخرجات الإعدادات الأولية..... 102
الشكل 6-2 المخرجات عند زيادة عدد الخطوات..... 103
الشكل 6-3 يوضح تحقيق الوكيل مكافأة أكبر عند استبدال تابع المكافأة..... 103
الشكل 6-4 وصول النموذج بالإعدادات النهائية للحالة المرغوبة..... 103
الشكل 6-5 النتائج الكمية والنوعية لاختبار السياسة الناتجة عن تجارب البناء والتدريب لحالة حساسات اللمس..... 104
الشكل 6-6 النتائج العينية لاختبارات المرحلة الثانية في حالة حساسات اللمس..... 105
الشكل 6-7 منحني المكافآت التراكمية لـ100 تجربة حققت 100% (أ) كنسبة نجاح في الإمساك، والأخرى (ب) حققت 91%..... 105
الشكل 6-8 نتائج تطبيق تعلم المناهج..... 106
الشكل 6-9 نتائج الاختبار على جسم غريب كليا..... 106
الشكل 6-10 نتائج التدريب على الإمساك بمطرقة بالإعدادات الأولية..... 107
الشكل 6-11 نتائج التدريب الكمية والاختبار العينية على إمساك المطرقة مع زيادة عدد الخطوات..... 107
الشكل 6-12 منحنيات التدريب للاختبار الأول في حالة حساسات القوة..... 108
الشكل 6-13 منحنيات تابع المكافأة التراكمية عبر 100 حلقة تجريب (أ)، و (ب) منحنيات فرق الضغط التراكمي عبر 100 حلقة..... 109
الشكل 6-14 قيمة القوة المطبقة في كل خطوة من أحد حلقات التجريب الفاشلة في عملية الإمساك..... 109
الشكل 6-15 منحنيات التدريب للاختبار الثاني من المرحلة الأولى، (أ) منحني المكافآت، (ب) منحني فرق الضغط التراكمي 110
الشكل 6-16 نتائج اختبار السياسة الناتجة عن الاختبار الثاني من مرحلة البناء والتدريب..... 110
الشكل 6-17 منحني فرق الضغط عبر الخطوات لثلاث حلقات من حلقات الاختبار للسياسة..... 110
الشكل 6-18 منحنيات التدريب للاختبار الثالث من مرحلة التدريب..... 111
الشكل 6-19 نتائج اختبار السياسة الناتجة عن الاختبار الثالث لمرحلة البناء..... 111
الشكل 6-20 منحنيات التدريب للاختبار الرابع من المرحلة الأولى..... 112

الشكل 6-21 نتائج اختبار السياسة الناتجة عن الاختبار الرابع. (أ). SMA للمكفأة التراكمي. (ب). SMA لفرق الضغط التراكمي. 112
الشكل 6-22 فرق الضغط المطبق في كل خطوة عند اختبار حلقة واحدة. 112
الشكل 6-23 منحنيات التدريب للاختبار الخامس من مرحلة البناء 113
الشكل 6-24 نتائج اختبار السياسة الناتجة عن الاختبار الخامس. 113
الشكل 6-25 فرق الضغط المطبق عند اختبار حلقة واحدة. 114
الشكل 6-26 تجربة الإزاحة. 115
الشكل 6-27 تجربة الإزاحة، ومع إزالة الطاولة طاولة. 115
الشكل 6-28 إعادة التجربة السابقة مع رفع العبوة 10 سم للأعلى 116
الشكل 6-29 التجربة مع فازة مزاحة 117
الشكل 6-30 زجاجة مع إزاحة أقل من 5 سم 117
الشكل 6-31 إزاحة أكبر من 5 سم، (أ) منحني المكافآت التراكمية، (ب) منحني فرق الضغط كمتوسط بالقيمة المطلقة. 117
الشكل 6-32 حلقتين مع إزاحة أقل من 5 (ناجحة) ويليها مع إزاحة أكبر من 5 سم (حلقة فاشلة). 118
الشكل 6-33 البرتقالة (قبضة ووزن وتوضع كف) مختلفين. 118
الشكل 6-34 تعلم المناهج لدفعة واحدة. 119
الشكل 6-35 تعلم المناهج على مرحلتين مرحلة 1 119
الشكل 6-36 تعلم مناهج مرحلة 2 120
الشكل 6-37 تجربة إمساك بيضة دجاجا بثلاث أصابع. 121
الشكل 6-38 التجربة لإمساك بيضة دجاج مزاحة 2 سم 122

فهرس الجداول

الجدول 1-2 مصفوفة الأداء (performance metrics) 23
الجدول 2-2 ملخص لبعض الدراسات السابقة الأخرى 26
الجدول 1-3 مقارنة بين مناهج التعليم الرئيسية للروبوتات (Kormushev et al, 2013). 42
الجدول 1-4 جدول لستة حالات مختلفة لدالة الهدف المقطوعة (Bick et al, 2021). 62
الجدول 2-4 مقارنة التعلم العميق بالتعلم التقليدي (Liu et al, 2021). 65
الجدول 1-5 خصائص الحاسوب الذي ستجرى عليه الدراسة. 72
الجدول 2-5 ناتج عملية التقطيع والتهيئة. 81
الجدول 3-5 عدد المفاصل ومواصفاتها التي تمت نمذجتها، (اللون البرتقالي لن يتم التحكم به لاحقا). 84
الجدول 4-5 المحددات الفائقة لخوارزمية PPO 92
الجدول 5-5 المحددات الفائقة لخوارزمية PPO بعد التحديث 97
الجدول 6-5 خصائص بيضة الدجاج المصممة (Joubrane et al, 2019). 99
الجدول 1-6 ملخص نتائج تجارب البناء والتدريب للسياسة في حالة حساسات اللمس. 103
الجدول 2-6 ملخص التجارب ومصفوفة الأداء لحالة حساسات اللمس Performance metrics 125
الجدول 3-6 ملخص نتائج الحالة الثانية ومصفوفة الأداء لحالة حساسات القوة Performance metrics 126
الجدول 4-6 مقارنة النتائج الحالية مع الدراسات السابقة. 130

فهرس المصطلحات

الرقم	المصطلح باللغة العربية	المصطلح باللغة الأجنبية	الاختصار
.1	الفاعل الناقد المتقدمة	Advantage Actor – Critic	A2C
.2	الذكاء الاصطناعي	Artificial Inelegant	AI
.3	المنطقة الأمامية من الثلم الجداري الداخلي	Anterior Intraparietal	AIP
.4	المنطقة الصدغية السفلى الأمامية	Anterior Inferior Temporal area	AIT
.5	الحزّ الجداري الداخلي الجانبي الأمامي	anterior lateral Intraparietal Sulcus	alPS
.6	القشرة الجدارية–القذالية العلوية الأمامية	anterior Superior Parieto-Occipital Cortex	aSPOC
.7	استنساخ السلوك	Behavioral Cloning	BC
.8	واجهة الدماغ الآلة	Brain–Machine Interface	BMI
.9	التعلم المؤتمت المستمر نهج فاعل ناقد	Continuous Actor-Critic Learning Automaton	CACLA
.10	ديناميكيات السوائل الحسابية	Computational Fluid Dynamics	CFD
.11	تعلم المناهج	Curriculum Learning	CL
.12	الرسغ السنعي	Carpometacarpal	CMC
.13	الشبكة العصبية الالتفافية	Convolution Neural Network	CNN
.14	تكرار السياسات المحافظ	Conservative Policy Iteration	CPI
.15	وحدة معالجة مركزية	Central Processing Unit	CPU
.16	الانحدار المتدرج للسياسة معزز بالعروض التوضيحية	Demonstration Augmented Policy Gradient	DAPG
.17	خوارزمية تدرج السياسة الحتمية العميقة	Deep Deterministic Policy Gradient	DDPG
.18	الانحدار المتدرج للسياسة الحتمية العميقة من العروض التوضيحية	Deep Deterministic Policy Gradient from Demonstrations	DDPGfD
.19	شبكة Q العميقة المزدوجة	Double Deep Q Network	DDQN
.20	التعلم العميق	Deep Learning	DL
.21	القشرة الجبهية الأمامية الظهرية الجانبية	Dorsolateral Prefrontal Cortex	DLPFC
.22	قشرة الفص الجبهي الأمامية الظهرية الجانبية	Dorsolateral Prefrontal Cortex	DLPFC
.23	البرمجة الديناميكية	Dynamic Programming	DP
.24	شبكة Q العميقة	Deep Q-Network	DQN
.25	سياسة الإمساك المعزز العميق	Deep Reinforcement Grasp Policy	DRGP
.26	التعلم المعزز العميق	Deep Reinforcement Learning	DRL
.27	القشرة الجبهية	Frontal	F5
.28	طريقة العناصر المحدودة	Finite Element Analysis	FEM
.29	وحدة معالجة الرسوميات	Graphics Processing Unit	GPU
.30	إعادة تشغيل الخبرة بأثر رجعي	Hindsight Experience Replay	HER
.31	التعلم المحاكي	Imitation Learning	IL
.32	التعلم المعزز العكسي	Inverse Reinforcement Learning	IRL
.33	معامل العقوبة التكيفي	Kullback–Leibler	KL
.34	مونت كارلو	Monte Carlo	MC
.35	السلاميات المشطية	Metacarpophalangeal	MCP
.36	عملية قرار ماركوف	Markov Decision Process	MDP
.37	تعلم الآلة	Machine Learning	ML

MPL	Modular Prosthetic Limb	الطرف الاصطناعي المعياري	.38
MPO	Maximum a posteriori Policy Optimisation	أمثلة السياسة بأقصى احتمال لاحق	.39
NPG	Natural Policy Gradient	تدرج السياسة الطبيعي	.40
ODE	Open Dynamics Engine	المحركات الفيزيائية	.41
PCA	Principal Component Analysis	تحليل المكونات الرئيسية	.42
PIP	Proximal Interphalangeal	بين السلاميات القريبة	.43
PMC	premotor cortex	القشرة قبل الحركية	.44
PPC	Posterior Parietal Cortex	القشرة الجدارية الخلفية	.45
PPC	Posterior Parietal Cortex	القشرة الجدارية الخلفية	.46
PPO	Proximal Policy Optimization	أمثلة السياسة القريبة	.47
pSPOC	posterior Superior Parieto-Occipital Cortex	القشرة الجدارية–القذالية العلوية الخلفية	.48
RA	Rapidly Adapting	التكيفية السريعة	.49
RFs	Receptive Fields	مجالات الاستقبال	.50
RGB- D	Red Green Blue – Depth	صور العمق الملونة	.51
RL	Reinforcement Learning	التعلم المعزز	.52
RTF	ratio of simulation time over real-world time	عامل الوقت الحقيقي	.53
SA	Slowly Adapting	التكيفية البطيئة	.54
SAC	Soft Actor Critic	خوارزمية الفاعل الناقد اللينة	.55
SGA	Stochastic Gradient Ascent	صعود التدرج العشوائي	.56
SGD	Stochastic Gradient Descent	نزول التدرج العشوائي	.57
SMA	Supplementary Motor	المنطقة الحركية التكميلية	.58
SMA	Simple Moving Average	المتوسط المتحرك	.59
TD	Temporal Difference	الفرق الزمني	.60
TD3	Twin Delayed Deep Deterministic Policy Gradient	انحدار السياسة الحتمي العميق ثنائية التقدير المؤجل	.61
TMC	Trapeziometacarpal	شبه السنعي المنحرف	.62
TRPO	Trust Region Policy Optimization	أمثلة السياسة في منطقة الثقة	.63
UR	Universal Robots	الروبوت العالمي	.64
V	Primary Visual Cortex	القشرة البصرية الثانوية	.65
VIP	Ventral Intraparietal area	المنطقة البطنية للثلم الجداري الداخلي	.66

الملخص

قادت الدراسات المكثفة في مجال الأطراف الاصطناعية العلوية لابتكار طرق تحكم متقدمة. ورغم ذلك، لا يزال من الصعب تصميم أطراف تعمل بمستويات عالية من البراعة والسيطرة تماثل قدرات اليد البشرية. شهد هذا المجال مؤخرا ثورة علمية بفضل التكامل التكنولوجي بين الذكاء الاصطناعي وعلوم الأطراف الاصطناعية، مما أدى إلى تطوير الأطراف الاصطناعية الذكية. تتناول هذه الرسالة توظيف التعلم المعزز العميق (وهو أحد فروع الذكاء الاصطناعي) بالإضافة لتعلم المناهج بهدف تمكين يد روبوتية ذات درجات حرية عالية من التعلم الذاتي للسلوك الحركي الأمثل عبر التفاعل المستمر مع البيئة. هذا من شأنه تعزيز الاستقلالية والذكاء الحركي مما يجعل أداء اليد الاصطناعية أقرب إلى الأداء البشري. وتكمن أهمية البحث في تطوير أطراف اصطناعية تفاعلية قادرة على التعامل مع بيئات ديناميكية متعددة يواجهها المستخدم، وتمكينها من أداء حركات الإمساك بكفاءة دون الحاجة إلى البرمجة التقليدية، وذلك اعتمادا فقط على المعلومات اللمسية.

أجريت تجارب هذا البحث في بيئة المحاكاة Webots، باستخدام خوارزمية أمثلة السياسة القريبة (Proximal Policy Optimization - PPO)، مع نموذج اليد الاصطناعية HANDY خماسية الأصابع ذات 11 درجة من الحرية، تم التحكم فعليا في 10 منها خلال التجارب. وأثناء عملية التدريب، سُجلت المشاهدات التي تمثل الحالة الآنية لليد وتشمل ردود الفعل اللمسية – والتي درست بحالتين: الاستشعار الثنائي لوجود التلامس أو عدمه، واستشعار القوة الذي يقيس قوة الضغط عند نقاط التلامس– بالإضافة لزوايا المفاصل. بناء على هذه الحالة يتخذ الوكيل الذكي لكل سياسة مدربة أحد إجراءين: إما زيادة زاوية المفصل بمقدار ثابت C أو الحفاظ على الوضع الحالي. ثم يقوم تابع المكافأة المصمم بتحفيز النموذج بناء على نوع الإجراء المتخذ، مما يسهم في أمثلة أداء السياسة مع مرور حلقات التدريب، تم تقييم عملية التدريب من خلال تتبع مجموعة من المقاييس الكمية والنوعية. ولاختبار كفاءة سياستي اتخاذ القرار الناتجتين، أجريت تجارب للتعميم واختبارات الصلابة، كما تم تحسين النتائج باستخدام مبدأ تعلم المناهج حيث حققت خلاله سياسات اتخاذ القرار تحسنا ملحوظا في نسب النجاح، وأخيرا أظهرت سياسة اتخاذ القرار نسبة نجاح 100% في حالة حساسات اللمس الثنائي، و96% في حالة حساسات القوة. وهو ما أظهر أن بنية النهج المقترح وهياكل المكافأة المطوّرة يتمتعان بقدرة عالية على التعميم، مما يجعلهما نهجا واعدا لتدريب اليد الاصطناعية على التفاعل مع مجموعة متنوعة من الأجسام.

الكلمات المفتاحية: الأطراف الاصطناعية العلوية، الذكاء الاصطناعي، التعلم المعزز العميق، أمثلة السياسة القريبة، محاكاة.

1. الفصل الأول: الإطار العام للبحث.

1.1. تمهيد:

الإنسان، هو النموذج الهندسي الأكثر تعقيداً وذكاءً الذي أوجده الله سبحانه وتعالى، ويشكل استنساخ بنيته التشريحية الدقيقة ووظائفه الحيوية تحديا كبيرا لمتخصصي الأطراف الاصطناعية وتقويم العظام. وقد اعتُبرت عملية التعويض عن النقص الوظيفي والجمالي الناتج عن البتر أو التشوهات في أطراف الكائنات مشكلة بحثية جوهرية استقطبت اهتماما واسعا من قبل الباحثين. حيث قادت الدراسات والبحوث المكثفة التي امتدت لعقود من الزمن إلى ابتكار تصاميم متعددة وطرائق تحكم متقدمة في الأطراف الاصطناعية. ورغم التقدم المحرز وبشكل خاص في مجال الأطراف الاصطناعية العلوية التي تتطلب آليات تحكم دقيقة ومعقدة لا يزال من الصعب حتى اليوم تصميم أطراف اصطناعية تحاكي براعة الأطراف البشرية من حيث الدقة والسيطرة.

مؤخرا، أدى التقدم السريع في مجالي الذكاء الاصطناعي والروبوتات إلى فتح آفاق واعدة جديدة تمنح الأمل لملايين الأشخاص من ذوي القدرات المختلفة الذين يواجهون تحديات وظيفية متنوعة. حيث أسفر الدمج التقني بين الذكاء الاصطناعي وهندسة الأطراف الاصطناعية (وخاصة علم التحكم في الأطراف الاصطناعية) إلى إحداث نقلة نوعية تمثلت في ظهور ما يعرف بالأطراف الاصطناعية الذكية (الأطراف الروبوتية). وعلى الرغم من هذا التقدم لا يزال تطبيق تقنيات الذكاء الاصطناعي في مجال الأطراف الاصطناعية وتقويم العظام في مراحله الأولى ولم يصل بعد إلى مراحل التطبيق الواسع أو التوظيف التجاري الكامل. إذ أن العديد من المشاريع القائمة في هذا السياق لاتزال في طور النماذج الأولية ولم تطرح في الأسواق إلا بشكل قليل جدا للاستخدام العام. وتعد حداثة المجال وقلة الخبرة والتكاليف المادية والحسابية المرتفعة أحد العوائق الرئيسية التي حدت من انتشار وإتاحة هذه الأطراف. وهنا تبرز الحاجة الملحة للبحث في هذا المجال بهدف تمكين أكبر عدد من المستفيدين من الوصول إلى تقنيات الأطراف الاصطناعية الذكية بأعلى مستويات من الجودة.

يستهدف البحث الحالي توظيف أحد مجالات الذكاء الاصطناعي الحديثة وهو التعلم المعزز (Reinforcement Learning - RL) إلى جانب التعلم العميق (Deep Learning - DL) أي التعلم المعزز العميق (Deep Reinforcement Learning – DRL) بغرض تطوير أساليب التحكم في كف يد اصطناعية روبوتية. إذ تتطلب الأطراف الروبوتية في الوقت الحاضر أن تكون تفاعلية قادرة على التكيف الذاتي لتتعامل مع حالات جديدة ومتنوعة. الأساليب القائمة على النموذج كطرق التحكم الكلاسيكية تتطلب ضبطا مملا للمحددات لكل مهمة في البيئة الثابتة وتكون غير قادرة على التعامل مع المهام الجديدة. لذا يُعد التعلم المعزز العميق أداة فعّالة لمعالجة البيئات الديناميكية بفضل قدرته على تعلم سياسات التحكم الجديدة بشكل مستقل من خلال نهج التجربة والخطأ وكميات ضخمة من البيانات والخبرة المتراكمة التي يجمعها آنيا. ونظرا لأن المهام الروبوتية وخاصة مهام الإمساك تعتمد على التفاعل المستمر بين الوكيل والبيئة فإن دمجها مع استراتيجيات التعلم المعزز يُعد أمرا محوريا بالغ الأهمية لتعزيز كفاءة

هذه المهام وتحقيق أداء أكثر مرونة وواقعية. حيث أن الهدف النهائي لدمج التعلم المعزز مع الروبوتات هو منح الروبوت القدرة على التعلم المستقل والتكيف وإنتاج المهام وأمثلتها في الشروط المتغيرة ديناميكيا وذلك على أساس الاستكشاف (Kormushev et al, 2013).

أثبتت الدراسات الحديثة أن خوارزميات التعلم المعزز تشكل اليوم عنصر أساسي في التطبيقات المعقدة التي تتراوح من ألعاب الفيديو إلى التحكم في الأطراف الاصطناعية والروبوتات. وكفرع من تعلم الآلة يستخدم التعلم المعزز تقنيات الذكاء الاصطناعي لتدريب الخوارزميات المعقدة في بيئة تفاعلية وقد تغلب بنجاح على التحكم التقليدي.

1.2. مشكلة البحث:

1. **الانتشار المرتفع لحالات البتر:** تشير التقديرات إلى أن 200 إلى 500 مليون عملية بتر تجرى سنويا على مستوى العالم (Nizamli, 2020). وفي الولايات المتحدة الأمريكية وحدها يعاني ما لا يقل عن 1.6 مليون شخص من بتر أحد أطرافهم ومن بين هذه الحالات 45% يعانون من بتر في الأطراف العليا ومن المتوقع أن تتضاعف هذه الإحصائيات بحلول العام 2050 (Ziegler et al, 2008). أما في سوريا، وفقا لمنظمة الصحة العالمية فقد سجلت نحو 30 ألف إصابة شهريا خلال السنوات الست الأولى من الثورة منها 1.5 مليون إصابة دائمة أدت 86 ألف حالة منها إلى بتر أحد الأطراف (Nizamli, 2020). كما أفاد بحث Tatar وزملاؤه أنه في عينة تضم 363 حالة بتر سورية كان أكثر من 25% منهم تقل أعمارهم عن 18 عام وما يقارب 90% كان قد سببها القصف، كما أظهر نسبا مرتفعة بحالات البتر السفلية والعلوية معا (Tatar et al, 2021). وبخلاف الصدمة الجسدية والعاطفية التي يتعرض لها الأشخاص الذين فقدوا أحد أطرافهم تظهر عواقب إضافية لا تقل أهمية كفقدان أحد أهم الوظائف الجسدية وتراجع مستوى استقلالية الفرد وقدرته على القيام بالمهام اليومية والقدرة على العمل والأنشطة الاجتماعية. إن معالجة هذه التحديات لا تقتصر على الجوانب الطبية وحسب بل تعود بفوائد جوهرية على جودة حياة الأفراد وتسهم في تعزيز استقرارهم النفسي والاجتماعي بما ينعكس إيجابيا على أسرهم ومجتمعهم من خلال تعزيز استقلاليتهم وتمكينهم من أداء المهام اليومية بكفاءة.
2. **القصور الوظيفي لأساليب وآليات التحكم التقليدية اليوم أمام متطلبات الحركة المعقدة:** توفر الأطراف الميكانيكية وظائف أساسية محدودة مثل القبض والبسط بينما تحاول الأطراف الإلكترونية المبرمجة تقليديا الاستفادة من المدخلات البشرية كالإشارات الدماغية أو العضلية لمحاكاة حركات أكثر دقة كالتحكم بإصبعين أو ثلاثة. ومع ذلك، تعاني هذه الطرق من ضعف الأداء والافتقار إلى الاستقرار وضرورة المعايرة المستمرة وعدم قدرتها على التكيف مع البيئات المتغيرة ما أدى إلى تخلي العديد من المستخدمين عنها (Sharif et al,

2020). وكذلك رفض الأطفال أطرافهم الاصطناعية بنسبة تصل إلى 50% (Redman, 2016) كما قرر واحد فقط من كل 4 مبتوري أطراف علوية استخدام الأطراف المُتحكم بها بالإشارات العضلية (James, 2021).

3. **الطبيعة المعقدة للمهارات الحركية الدقيقة لليد كالإمساك والمعالجة اليدوية:** تعد عملية استعادة وظائف اليد البشرية بما في ذلك القبضات المتنوعة والمهارات الحركية الدقيقة التي تتطلب تحكما منسقا ومعقدا تحديا كبيرا في تصميم الأطراف الاصطناعية، ويرجع ذلك أساسا إلى ارتفاع عدد درجات الحرية والتي تملك اليد البشرية 27 منها، بالإضافة لتعقيد البنى العصبية والميكانيكية المرتبطة بها (Rajeswaran et al, 2018) (Kepes, 2017).

4. **محدودية أساليب تعلم الآلة الأخرى أمام التعلم المعزز:** تُمكن تقنية تعلم الآلة من التحول في حل المشكلات من النهج التحليلي إلى النهج القوي القائم على البيانات. وعلى الرغم من التقدم الكبير الذي يحرزه الباحثون في تقنيات التحكم بالأطراف الاصطناعية باستخدام أساليب تعلم الآلة (الموجه وغير الموجه) إلا أن هناك بعض التحديات التي تواجه أنظمة الإمساك الآلي باستخدام هذه التقنيات، مثلا تحتاج أساليب التعلم الموجه إلى كم كبير من البيانات المجمعة مسبقا والمصنفة والتي يصعب الحصول عليها في مهام الأيدي الروبوتية ذات درجات الحرية العالية، كما يتطلب التعلم غير الموجه الحصول على كم كبير من البيانات واستخلاص خصائصها وسماتها للتعلم دون أي توجيه مسبق، وهنا تظهر ميزة التعلم المعزز العميق فالتعلم المعزز لا يشبه تقنيات تعلم الآلة السابقة لأنه غير مدرب على البيانات المصنفة كما هو الحال في التعلم الموجه بل يجب أن يجد أفضل الحلول من خلال الاكتشاف عن طريق التجربة والخطأ حتى يكتشف كيفية تحقيق الهدف ولكن على عكس التعلم غير الموجه لا يتعين عليه العثور على السمات الموجودة في بيانات التدريب دون أي دليل بطريقة غير موجهة وبدلاً من ذلك يتم تشغيله بواسطة إشارة مكافأة. ما يجعل هذا النموذج مرن للغاية وهناك العديد من الخوارزميات والتطبيقات المختلفة لمختلف مسائل RL (Liu et al, 2021) (Mellatshahi, 2020).

5. **حداثة مجال التعلم المعزز العميق وقيود خوارزمياته:** إذ تحتاج خوارزمياته لخبرة عملية عالية لتطبيقها في مجال الروبوتات وتعد قيودها التقنية وحداثتها النسبية ونقص الخبرات التطبيقية من أهم الأسباب التي حدت من تطبيقها اليوم وهو ما يتطلب تكثيف الجهود البحثية وتعزيز الدراسات المتخصصة في هذا المجال الواعد.

ومن هنا ظهرت الحاجة لأهمية ابتكار أساليب تحكم متقدمة تخلق نقلة نوعية في عالم التحكم بالأطراف.

1.3. فجوة البحث:

تعد الأطراف الاصطناعية التجارية التقليدية المتاحة اليوم أطراف محدودة الذكاء تقتصر في أداءها على أوامر المستخدم المباشرة وتفتقر إلى فهم نوايا المستخدم أو استيعاب البيئة المحيطة والسياق الذي تُستخدم فيه، ونتيجة لذلك تقتصر استجابتها على تنفيذ الأوامر الصريحة الصادرة عن المستخدم، وبهذا تكون قدراتها محصورة ضمن

نطاق واجهة التحكم فقط حيث تعجز الذراع الاصطناعية عن تنفيذ أي مهام تتجاوز ما يُمليه عليها المستخدم بشكل مباشر. وتبرز في هذا السياق إحدى التحديات المحورية، وهي الفجوة الواسعة بين عدد الوظائف المحدود التي تتيحها هذه الأطراف من جهة، وتنوع الأوامر والبيئات التشغيلية التي تُستخدم فيها من جهة أخرى. إذ تعد اليد البشرية نظاما قويا للاستشعار والعمل ضمن البيئات المختلفة، وبهدف تطوير أطراف تحاكي قدراتها الدقيقة يجب العمل على تطوير أطراف ذكية تفاعلية، وأحد السبل الواعدة التي يحقق فيها الباحثون لتحقيق ذلك، هو دمج أساليب الذكاء الاصطناعي وتعلم الآلة العميق المتنوعة ضمن آليات التحكم في الأطراف الاصطناعية في المنطقة التي يتوقف عندها التحكم التقليدي. فيصبح من الممكن تطوير أطراف قادرة على توقع احتياج المبتور ضمن بيئة التفاعل واتخاذ الإجراء المناسب لمساعدته، وبالتالي تجاوز فجوة التحكم التقليدي التي تفرضها واجهة التحكم. ويمكن توضيح المشكلة بشكل أكبر من خلال تصور شخص بترت ذراعه على مستوى الساعد أو حتى فوق الكوع، ويستخدم طرفا اصطناعيا تقليديا مقاد بالإشارات العضلية، إن كل ما يمكن للمستخدم القيام به من أجل إمساك جسم ما موضوع على الطاولة، هو توجيه الطرف نحو الجسم المراد إمساكه، ومن ثم الإشارة إلى قبضة عامة للإمساك أو التحرير. لكن ما الذي سيحدث بعد ذلك يمثل تحديا حقيقيا، إذ أن العديد من العضلات والأعصاب والأوتار التي كانت مسؤولة عن التحكم الدقيق بحركات الأصابع قد فقدت مما أدى إلى غياب القدرة على تحديد النوايا الحركية الدقيقة للمستخدم كالثني أو البسط في الأَصابع الاصطناعية. وهنا يتولى الباحثون اليوم في مجال الأطراف الاصطناعية المسؤولية في العمل على تطوير أطراف ذكية تفاعلية يمكنها الاكتشاف الذاتي للقبضات المثالية. حيث يتم العمل على تطوير نماذج تعلم آلي للتحكم في الأيدي الاصطناعية يمكنها ملاحظة الأوامر العضلية أو العصبية للمستخدمين فضلا عن مراقبة حالة الإمساك والبيئة المحيطة لتوليد أفضل قبضة.

وبالإضافة لهذه الفجوة في التحكم وعلى الرغم من التقدم في تقنيات التحكم بالأطراف الاصطناعية والذكية بشكل خاص، إلا أن قلة الأبحاث التي تعتمد على التغذية الراجعة اللمسية بدلا من البصرية، رغم أهميتها في تقريب الأداء الاصطناعي من الأداء الحيوي، وندرة الدراسات التي تتناول تدريب يد روبوتية كاملة بخمس أصابع وبدرجات حرية عالية، تظهر وجود فجوة واضحة في الأدبيات العلمية. إضافة إلى أن اعتماد خوارزميات التعلم المعزز العميق في بيئات بدءاً من الصفر (دون بيانات مسبقة) باستخدام نهج التجربة والخطأ فقط، لا يزال مجالا غير مستغل بشكل كاف، مما يدعو إلى استكشاف هذا التوجه الجديد.

1.4. أهداف البحث:

يهدف البحث الحالي إلى توظيف التعلم المعزز العميق المعتمد على ردود الفعل اللمسية بنوعيها (الثنائية أو القوة) وزوايا المفاصل لتدريب كف روبوتية خماسية الأصابع ذات إحدى عشر درجة من الحرية (يتم التحكم فعليا في عشرة منها)، على تعلم سلوكيات الإمساك الأمثل من خلال التفاعل المستمر مع البيئة ضمن المحاكاة. ويركز

البحث على تقييم فعالية هذا النهج في أمثلة استراتيجيات التحكم بالأطراف الاصطناعية ذات درجات الحرية العالية، وتعزيز الذكاء الحركي والاستقلالية في أداء المهام، مما يجعلها أقرب في أدائها إلى اليد البشرية.

1.5. الدافع وراء البحث:

تركز معظم الدراسات في مجال التعلم المعزز وأيدي الأطراف الاصطناعية على المعلومات البصرية فقط كتغذية عكسية من البيئة المحيطة وعلى الرغم من أن الرؤية تعطي ردود فعل مهمة أثناء تحريك الأيدي نحو الهدف إلا أنه عمليا لا يُحتاج إليها كثيرا فبمجرد تحديد الهدف وموقعه من خلال الإدراك البصري يمكن تحريك الأيدي إلى الموقع والتقاط الأشياء بأعين مغلقة وبدقة مذهلة، فنحن البشر نستخدم ذكاءنا اللمسي الذي يساعد بالشعور بوزن وشكل الكائن والجمع بين هذه الخصائص وغيرها لتعديل حركة وشكل قبضة اليد بما يتوافق مع ذلك بهدف الإمساك أو التحكم الحركي الدقيق. يلعب الاستشعار اللمسي بنوعيه: ((Pressure (Force)/Touch (Binary) دور مهم في عمليات الإمساك والتحكم الحركي الدقيق في اليد البشرية وكذلك في أيدي الروبوتات، فهو يوفر ملاحظات حول اتصالاته الفيزيائية والقدرة على التفاعل الدقيق مع البيئة والتي تعزز استقرار القبضة وتفتقر إليها وسائل الاستشعار الأخرى مثل أجهزة استشعار الرؤية. وعليه وبسبب الإمكانات الكبيرة التي يتيحها الاستشعار اللمسي في عمليات الإمساك ونتيجة قلة أبحاث التعلم المعزز من الصفر التي تعتمد على التغذية الراجعة اللمسية والتي تتناول تدريب يد روبوتية بدرجات حرية مرتفعة، ظهر الدافع للبحث، بتطوير أساليب التحكم في الأطراف الاصطناعية.

1.6. فرضية البحث:

في البحث نريد برهان فيما إذا كانت خوارزمية DRL المقادة بردود الفعل اللمسية تقنية مفيدة يمكن استخدامها في مجال التحكم في الأطراف الاصطناعية لتعليم نموذج كف اصطناعي ذو درجات حرية عالية مهمة الإمساك أم لا.

1.7. أهمية البحث:

تتمثل أهمية هذا البحث في تطوير طرف اصطناعي ذكي ذي درجات حرية عالية قادر على التفاعل والتكيف مع البيئات الديناميكية المتنوعة باستخدام التعلم المعزز العميق المعتمد على اللمس، حيث يُعد نهج التعلم المعزز من أكثر الأساليب الواعدة لبناء روبوتات ذكية قادرة على التكيف مع المتغيرات والاضطرابات دون الحاجة إلى إعادة برمجة تقليدية أو إعداد معقد للمحددات أو للنماذج (Franceschetti et al, 2020) (Sekkat et al, 2023). كما يتغلب على تحديات التعلم التي تفرضها مساحات الحالة ذات الأبعاد العالية، وهو ما يميز الروبوتات الحديثة متعددة الحركات.

من المزايا الإضافية لهذا النهج:

1. القدرة على التعلم من الصفر دون الحاجة لقواعد بيانات ضخمة مجمعة مسبقا، وهو ما يمثل تحديا بالنسبة لخوارزميات تعلم الآلة التقليدية وخاصة عند تطبيقها في مجال التحكم بالأطراف. فالتعلم المعزز غير مدرب على البيانات المصنفة كما هو الحال في التعلم الموجه بل يجب أن يجد أفضل الحلول من خلال الاكتشاف عن طريق التجربة والخطأ، ولكن على عكس التعلم غير الموجه لا يتعين عليه العثور على الخصائص الموجودة في بيانات التدريب دون أي دليل بطريقة غير موجه وبدلا من ذلك يتم توجيهه بواسطة المكافأة (Kepes, 2017).

2. تجاوز التحديات المرتبطة بالمدخلات البصرية مثل سوء الإضاءة وانسداد الرؤية ودقة الكاميرا التي تؤثر على ثبات الأداء، وذلك من خلال الاعتماد بشكل كامل على المدخلات اللمسية وإلغاء المدخلات البصرية. وكما هو معروف إن المعلومات اللمسية تساعد في تنوع التفاعل المطلوب للمعالجة اليدوية حيث إنها تمكن من استيعاب الأشياء بقوة بغض النظر عن شكلها أو خصائصها المادية. ومع ذلك، فإن دمج الإشارات اللمسية عالية الأبعاد في وحدة تحكم ومعالجة داخلية تعد قضية معقدة بالفعل وليست بالأمر السهل (Veiga et al, 2020).

3. دعم تعلم قبضات أكثر تنوع وواقعية عبر التفاعل المباشر مع البيئة، فعلى الرغم من أننا قادرون –وإن كان ذلك بصعوبة في بعض الحالات– على بناء وبرمجة حركات ومسارات الأيدي الاصطناعية يدويا لتؤدي مهام خارقة بأداء عالي الدقة وتتفوق به إلا أن هذا النهج ومع تنوع حركات اليد البشرية وتعقيدها وزيادة عدد درجات الحرية لكل مفصل سيتطلب جهود كبيرة، كما ستفشل اليد عند وضعها بأي سيناريو حركي جديد في أداء المهمة بطريقة صحيحة فيصبح من غير المجدي للمهندسين تطوير مهام معقدة في الأطراف الحديثة، تعرف الأيدي المبرمجة يدويا بالأيدي العمياء وهي التي لا تستخدم أي ردود فعل من البيئة المحيطة (Kepes, 2017). لذا تكمن أهمية البحث الرئيسية في أهمية الحصول على أطراف اصطناعية تفاعلية يمكنها تحقيق الحركات المعقدة للإمساك والتلاعب بالأشياء المتواجدة في الأيدي البشرية بسهولة ذلك باستخدام المنهج التجريبي الذي يقوم عليه DRL، والحصول على أطراف تفاعلية مما يلغي حاجة المبتور لاستبدال الطرف في البيئات المختلفة.

4. السعي لتحقيق ما تهدف له معظم الدراسات الحديثة بجعل وظائف الإمساك والتحكم اليدوي الدقيق أكثر عمومية بحيث تُدرب الأيدي الاصطناعية على تشكيل حركات الإمساك الضرورية والتخطيط التلقائي لإكمال المهمة أو مجموعة مهام دون الحاجة للبرمجة اليدوية المسبقة، أو حتى البيانات المجمعة مسبقا، ذلك باستخدام المنهج التجريبي الذي يستخدم البيانات التجريبية لاختيار التموضع الصحيح لأداء المهمة المحددة، وهو ما تعتمد عليه طريقة التعلم المعزز العميق الذي يعتمد مفهوم التجربة والخطأ في التعلم.

1.8. مساهمات البحث:

تتلخص المساهمات الأساسية لهذا البحث على النحو التالي:

1. الدراسة النظرية لآليات التحكم في اليد البشرية ولبُتر الأطراف العلوية، بهدف فهم كيفية محاكاتها ضمن DRL.
2. مراجعة طرق التحكم في الأطراف العلوية المتاحة حاليا وتاريخ استخدام AI مع الأطراف الاصطناعية العلوية.
3. الإلمام بالمفاهيم الأساسية والأساليب المتبعة في كل من التعلم المعزز والتعلم العميق وتعلم المناهج.
4. مراجعة الأبحاث والأوراق البحثية المتوفرة ذات الصلة، بهدف فهم DRL وكيفية تطبيقه على الإمساك متعدد الأصابع والقائم على اللمس بشكل خاص، مما يتيح القدرة على تقييم النتائج ومقارنتها بالأبحاث الحالية.
5. البحث في كيفية ومبدأ تطبيق خوارزمية PPO في حالة الإمساك متعدد الأصابع القائم على اللمس، رغم وجود أبحاث مشابه قائمة على الرؤية بهدف مقارنة النتائج.
6. تحديد فوائد تطبيق تعلم المناهج (Curriculum Learning - CL)، لبناء حلول تدريجية لمهام الإمساك المعقدة.
7. الاستفادة بعملية التدريب من المعلومات اللمسية لحساسات لمس موضوعة على اليد بدلا من المعلومات البصرية واختبار مدى فعاليتها في التدريب.
8. تعلم إجراءات التحكم في الطرف الاصطناعي الروبوتي بشكل مستقل لتنفيذ مهام جديدة، دون الحاجة إلى أي بيانات حقيقية و/أو عروض بشرية، ودون معرفة سابقة بديناميكيات النظام.
9. دراسة موضوع اللمس Touch وكذلك قوة القبضة Force، ومدى تأثيرها على إنتاج قبضات فعالة، ومقارنتها بالدراسات السابقة المشابهة أو القائمة على الرؤية.
10. جمع النظريات مع الممارسة العملية وكذلك استكشاف آليات التنفيذ العملية لأساليب التعلم المعزز العميق في تدريب الروبوتات والتي تحتاج بيئات محاكاة ولغات برمجية متقدمة.

1.9. تحديات البحث:

تم تصميم الأيدي الروبوتية متعددة الأصابع (والتي يشار إليها عادةً بالأيدي البارعة) لتحقيق مهارات حركية دقيقة عالية تصل لمستوى الإنسان أو شبيهة بالإنسان للروبوتات أو كأطراف اصطناعية. رغم تاريخ البحث الكبير في هذا المجال، لا تزال هناك تحديات كبيرة لتصميمها والتحكم فيها بشكل فعال نظرا لأبعادها العالية، وأوضاع التفاعل المتغيرة بشكل متكرر ومتطلبات التعميم على المهام المختلفة.

وعليه إن تطبيق DRL على الروبوتات أمر صعب رغم إمكاناته العالية، وتتبع الصعوبة من عدة نقاط تتمثل بـ:

1. كفاءة العينة: قد تكون بيانات التفاعل في العالم الحقيقي مكلفة وتستغرق وقتا طويلا لجمعها أنيا للخوارزميات.
2. اختيار بيئات المحاكاة المناسبة وإعداداتها، هناك الكثير من بيئات المحاكاة والمكتبات القابلة للاستخدام في ظل أنظمة تشغيل متنوعة لها مزايا وعيوب مختلفة. دون تحقيق جيد حول البيئات المتوفرة وانتقاء الأنسب قبل الاستخدام قد يتسبب ذلك لاحقا في حدوث مشكلات عديدة، في دقة النماذج ونشرها.

3. فجوة نقل المحاكاة إلى الواقع: تسد المحاكاة هذه الفجوة بين خوارزميات التعلم المعزز ومشاكل تدريب الروبوتات في العالم الحقيقي. إلا أنه بسبب الديناميكيات والضوضاء الموجودة في العالم الحقيقي، وتبسيطات المحاكاة قد لا تنقل تقنيات المحاكاة إلى إعدادات العالم الحقيقي بكفاءة، ويعتمد ذلك بشكل كبير على دقة المحاكاة.
4. السلامة والمتانة والتعميم: يجب أن تعمل الروبوتات المدربة بأمان وموثوقية في إعدادات ديناميكية وغير متوقعة.
5. غالبا ما تكون النهايات في مثل هذه الأبحاث مفتوحة، فبسبب الأشكال الكثيرة من مهام الإمساك والإعدادات ينتج عدد كبير من التجارب الممكنة مما يجعل من الصعب تحديد ما يجب التركيز عليه بشكل محدد.
6. لاتزال أحدث النتائج حتى اليوم في مرحلة مبكرة مقارنة بعملية الإمساك لدى اليد البشرية –على الرغم من أن التحكم اليدوي الدقيق الآلي متعدد الأصابع بالأشياء يعد مجال بحث واسع– وذلك بسبب لعنة الأبعاد إذ لا يزال التدريب على نماذج اليد ذات درجات الحرية الكبيرة محدود للغاية.
7. اختيار الخوارزمية المناسبة والحساسية الكبيرة لخوارزميات DRL للإعداد الأولي وضبط المحددات الفائقة، فهناك عدد كبير من المحددات الفائقة والإعدادات لإطار عمل DRL والمحاكاة التي يمكن أن تؤثر بشكل كبير على النتائج وتحددها. ورغم أن الهدف هو جعل النماذج عامة قدر الإمكان فإن العثور على المحددات الفائقة لا يزال تحدي رئيسي لخوارزميات تعلم الآلة الحديثة.
8. المفاضلة بين الاستغلال والاستكشاف، من خلال الاستكشاف يمكن للروبوت الحصول على مزيد من المعلومات حول البيئة لكن السلوك العشوائي المتخذ في هذه الحالات قد لا يحصل على مكافآت في المهام ذات المكافآت المتفرقة وتعد المكافآت نقطة مهمة لجعل النماذج تتقارب مع الهدف. من ناحية أخرى يستفيد الاستغلال من المعرفة حول البيئة لاتخاذ أفضل قرار لكن نقص المعلومات قد يؤدي إلى الوقوع بمشكلة الحل الأمثل المحلي لذلك، يجب أن يُحدد عاملان مهمان: كيفية الاستكشاف بكفاءة وفعالية، ومتى يجب الانتقال إلى الاستغلال.
9. متطلبات الأجهزة والوقت، على الرغم من أن خوارزميات تعلم الآلة القائمة على البيانات أصبحت أكثر كفاءة بشكل ملحوظ في السنوات الأخيرة إلا أنها تستغرق وقتا طويلا للتدريب من بضع ساعات إلى أكثر من أسبوع في بعض الحالات. بشكل عام، أفضل حل لهذه المشكلة هو تنفيذ الإصدارات الموزعة من هذه الخوارزميات والتي تمثل تحديا خاصة عندما تكون الموارد الحسابية محدودة.
10. تعتبر المواضع الأولية والضبط الأولي للبيئات كزوايا الإصبع وموضع الكف والكائن أمرا ضروريا أيضا لزيادة فرص الحصول على نتائج جيدة.
11. اختيار القابض المناسب، له دور كبير في النتائج العملية.

يمكن التغلب على هذه القيود بتطوير أجهزة محاكاة أكثر تقدما، وأمثلة الخوارزميات، وتطوير تصاميم الأيدي.

1.10. هيكل الرسالة:

نظمت الفصول القادمة لهذه الرسالة على النحو التالي:

1. **الفصل الأول** (الإطار العام للبحث)، تضمن مقدمة تمهيدية للموضوع المدروس، ثم عرض مشكلة البحث وأهميته مبينا الدافع والفرضية الأساسية للبحث، وكما وضح التحديثات والمساهمات التي قدمها البحث.
2. ثم عُرضت مراجعة لعدد من الدراسات السابقة ذات الصلة والدراسات التي تناولت تطبيق التعلم المعزز في مجال التحكم بالأطراف والأيدي الروبوتية، وذلك بهدف فهم مستوى التقدم في هذا المجال وتحديد الإضافات والمساهمات التي يقدمها هذا البحث، وذلك في **الفصل الثاني** (الدراسات السابقة ذات الصلة).
3. يلي ذلك دراسة نظرية لآليات التحكم في اليد البشرية ودراسة فسيولوجية وتشريحية لليد وعمليات البتر وكذلك مراجعة لأساليب التحكم في الأطراف الاصطناعية، في **الفصل الثالث** (اليد البشرية والأطراف الاصطناعية: دراسة بيولوجية ووظيفية لآليات التحكم والإمساك).
4. تم توضيح جملة من المفاهيم والأسس النظرية التي اعتُمد عليها في **الفصل الرابع** (الأسس نظرية للبحث).
5. في **الفصل الخامس** (منهجية العمل)، كما يوحي اسمه عرض هذا الفصل منهجية البحث وخطواته بشكل موسع ومنظم، بدأ من تحديد الأدوات اللازمة لإنجاز البحث بالاعتماد على توصيات الدراسات المرجعية، ثم تهيئة بيئة التدريب، وانتهاء بجزء التعلم المعزز العميق وتعلم المناهج.
6. يعرض **الفصل السادس** (النتائج ومناقشتها) النتائج التي تم الحصول عليها ويناقشها ويجري مقارنة بين نتائج البحث والدراسات السابقة ذات الصلة.
7. تُختتم الرسالة بعرض الاستنتاجات التي تم التوصل إليها، والتطلعات المستقبلية وذلك ضمن **الفصل السابع** (الاستنتاجات والآفاق المستقبلية).

1.11. الخاتمة:

يعد استخدام DRL في مجال التحكم في الأطراف الاصطناعية ولا سيما العلوية موضوع بحثي ناشئ لايزال في مراحله الأولى ويعزى ذلك إلى حداثة هذا التخصص من جهة وتعقيد المهام الحركية المطلوبة وارتفاع عدد درجات الحرية خاصة عند محاولة محاكاة اليد البشرية. وتضاف إلى ذلك قيود مرتبطة بالخوارزميات المستخدمة وآليات تنفيذها ما يشكل تحديا إضافيا أمام التقدم في هذا المجال. ومع ذلك، يشهد المجال تطورا متسارعا يتمثل في اقتراح خوارزميات جديدة وفعالة تسعى لتجاوز التحديات. وبحسب ما تظهره الدراسات الحديثة فإن الدمج بين تقنيات AI وخوارزميات التحكم التكيفية ينبئ بإمكانية الوصول إلى أنظمة تحكم أكثر تطورا ومرونة، قادرة على تنفيذ حركات دقيقة ومعقدة بصورة أسهل وأكثر كفاءة وذلك على خلاف الأساليب التقليدية محدودة القدرات والصعبة التنفيذ للحركات المعقدة.

2. الفصل الثاني: الدراسات السابقة ذات الصلة

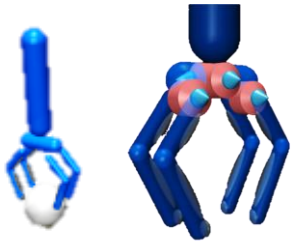
2.1. تمهيد:

اكتسبت مناهج تعلم الإمساك القائمة على البيانات في العقدين الأخيرين بشكل عام شهرة متزايدة في مجال الروبوتات. إذ تُعد مهمة الإمساك الآلي من المهام المعقدة والمثيرة التي تستقطب اهتمام الباحثين، مما دفع العديد من المراجعات والدراسات لتسليط الضوء على مختلف الأساليب والتقنيات المستخدمة في هذا السياق. وفي السنوات الأخيرة، لاقت سياسات الإمساك الروبوتية المدعومة بالتعلم المعزز العميق اهتماما كبيرا باعتبارها نهجا واعدا. في هذا الفصل، تُعرض أحدث المراجعات حول تطبيق التعلم المعزز العميق في مهمة الإمساك. حيث تُقدم نظرة عامة شاملة على الأبحاث الحالية في مجال الإمساك الروبوتي، مع التركيز على الدراسات التي تناولت تقنيات الإمساك باستخدام آليات DRL المختلفة. تم تصنيف الدراسات المعروضة في هذا الفصل إلى خمس فئات رئيسية، وذلك استنادا إلى نوع التغذية الراجعة المستخدمة وطبيعة وتعقيد البيئية التجريبية، بالإضافة لدراسات التعلم من العروض التوضيحية بهدف مقارنة نتائجها لاحقا مع دراسات التعلم بدأ من الصفر، وشملت هذه الفئات دراسات تأخذ معلومات لمسية، وأخرى بصرية، ودراسات تأخذ قيم عددية كالمسافات الإقليدية والإحداثيات، ودراسات مع بيئات معقدة.

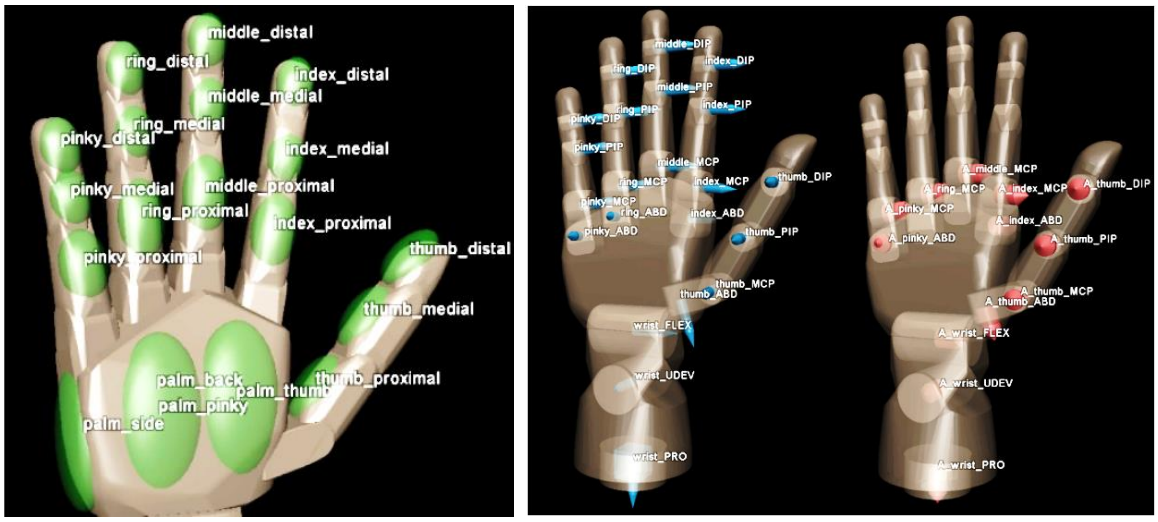
2.2. دراسات تعتمد التغذية العكسية اللمسية:

أجرى Kepes عام 2017 دراسة بهدف تعليم يد روبوتية أداء مهام الإمساك الدقيقة باستخدام ردود الفعل اللمسية فقط ضمن بيئة محاكاة. اعتمدت الدراسة على محاكي MuJoCo وتم تدريب نموذجين: الأول ويتألف من ذراع قابض بمخلبين ورسغ يملك 7 درجات حرية (استخدم منها 3 فقط) كما هو مبين في الشكل (1−2)، والثاني يظهر في الشكل (2−2) ويتألف من نموذج اليد الاصطناعية المعيارية (Modular Prosthetic Limb - MPL) والتي تمتلك 25 درجة حرية (استخدم 7 منها فقط). استخدمت خوارزمية (Double Deep Q Network – DDQN) لتحقيق الإمساك ورفع أجسام بسيطة موضحة في الشكل (3−2)، حيث يتم قراءة حساسات اللمس فقط أو زوايا المفاصل وحساسات اللمس ثم تُعدل زاوية كل مفصل بالزيادة أو النقصان أو تركها كما هي، وتُعيّن المكافأة تبعا لارتفاع الكائن الملتقط. أظهرت النتائج أن الذراع الأول حقق معدل نجاح قدره 40% بعد 2500 حلقة تدريب لكل منها 250 خطوة عند رفع كرة من ارتفاع ثابت مع حساسات اللمس فقط في حين تحسن الأداء بشكل ملحوظ مع تضمين زوايا المفاصل، ثم عند التدريب من الصفر على رفع الكرة مع ارتفاع بداية متفاوت فشل الوكيل لذا قام الباحث باستخدام تعلم المناهج حيث تابع تدريب النموذج الأول المدرب مسبقا 2000 حلقة إضافية فتعلم الوكيل رفع الكرة من ارتفاعات متفاوتة، وأظهر نقل النموذج المدرب إلى اليد الثانية ذات درجات الحرية الأعلى (بتفعيل الإبهام والسبابة فقط) ورفع عدد الخطوات إلى 500 خطوة، قدرة على التكيف السريع مع تغيرات البيئة (نموذج اليد، والارتفاعات العشوائية) ما دل على إمكانية التعميم. بذلك، أظهرت الدراسة دور أسلوب تعلم المناهج في أمثلة أداء النماذج المدربة مسبقا. لاحقا لذلك جُرّبت يدويا قيم مختلفة للمعلمة الفائقة ϵ وقد حققت أحد الإعدادات معدل

نجاح 100% مع 2500 حلقة زمنية بدلا من 6000 – 8000 حلقة مع قيم ϵ أخرى أي بزمن تعلم أقصر بنسبة 65% ما يظهر أهمية ضبط المحددات الفائقة، بينما فشل النموذج لاحقا في تكرار النجاح ذاته عند زيادة التعقيد إلى ثلاث أصابع (Kepes, 2017). كانت النتائج مشجعة لاستكشاف المزيد من الأبحاث حول فعالية حساسات اللمس وزوايا المفاصل كما بينت الدراسة أهمية أسلوب تعلم المناهج، إلا أن أحد الأسباب المحتملة لحد عملية زيادة التعقيد قد يعود لنوع الخوارزمية حيث يعرف أن خوارزميات تعلم **Q** لها صعوبات في التنفيذ وضبط المحددات، وكان من الجيد بحسب ما أوصت الدراسة استعراض خوارزميات تدرج السياسة المعروفة بسهولة تنفيذها وضبط محدداتها.

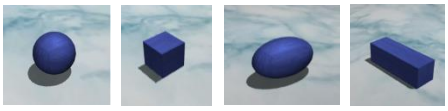


الشكل 1–2 نموذج اليد الأول يمثل اللون الأحمر مواضع المحركات واللون الأزرق المفاصل واللون الرمادي حساسات اللمس (Kepes, 2017).





الشكل 2–2 نموذج اليد الثاني المستخدم يمثل اللون الأحمر موضع المحركات والأزرق المفاصل والأخضر حساسات اللمس (Kepes, 2017).



الشكل 3–2 الأجسام التي تم التعامل معها أثناء التدريب والاختبار (Kepes, 2017).

في دراسة حديثة قام Zhang وزملاؤه عام 2025 بدراسة تأثير الإدخال اللمسي على أداء الإمساك لروبوت ثنائي الأصابع ذو درجة حرية واحدة مثبت على ذراع Franka Panda، باستخدام محاكي IssacGym. اختُبر الأداء في حالات رؤية مثالية وغير مثالية مع مستشعرات لمسية من أنواع متعددة (ثنائية، مطال القوة، متجه القوة)، وبتطبيق خوارزميتي (Soft Actor Critic – SAC) و(Maximum a posteriori Policy Optimisation - MPO). أثناء التدريب على مهمة الوصول والإمساك لثلاث نماذج أولية هي: علبة مشروبات معدنية وصندوق وموزة بلاستيكية يقرأ الوكيل بيانات الكاميرا والمستشعر اللمسي ويتحكم بمواقع المفاصل. بعد التدريب تم تقييم الأداء ثم نفذت تجارب النقل للواقع باستخدام نفس الإعدادات للتحقق من النتائج على الروبوت الحقيقي. أظهرت النتائج أن الإدخال اللمسي لا يؤثر كثيرا في ظروف الرؤية المثالية وكان معدل نجاح الإمساك 60% تقريبا، ويستنتج من ذلك أن الرؤية المثالية تكفي وحدها ولا تحتاج دعما لمسيا، إلا أنه من الصعب جدا أن تتحقق في الواقع وبالتالي تمت المحاكاة مرة أخرى مع وجود الضوضاء فوُجد أن الاستشعار اللمسي يحسن الأداء بشكل واضح في بيئات الرؤية المشوشة وبشكل خاص مع خوارزمية MPO حيث يتراجع الأداء مع رؤية مشوشة دون مستشعر. ويُستنتج من ذلك أن الاستشعار اللمسي يقلل تأثير عيوب الرؤية وكان ذلك أوضح مع حساسات متجه القوة. كما تبين أن الاستشعار اللمسي يعزز من تعميم السياسة على أجسام جديدة وغير مألوفة. في تجارب نقل السياسة من المحاكاة للواقع عند نقل السياسة بدون اللمس فشلت السياسة (حققت نسبة نجاح 0%)، ومع مستشعر ثنائي حققت نسبة نجاح تصل حتى 55% ومع مستشعر مطال القوة حققت حتى 66% ومع مستشعر متجه القوة حققت حتى 70%، كما أن تأثير نوع خوارزمية التعلم كان واضحا، حيث أظهرت MPO أداء أكثر استقرارا مقارنة بـ SAC وذلك لأن تحديث السياسة في MPO أكثر تحفظا (تحديث تدريجي وآمن) ما يجعلها أقل حساسية للتقلبات (Zhang et al, 2025). أكدت هذه الدراسة أهمية المستشعرات اللمسية في مهام الإمساك الروبوتي، إلا أنها لم تتعامل إلا مع درجة حرية واحدة.

في دراسة عام 2022 اقترح Matak نهجا ينشئ قبضات دقيقة بالاعتماد على مدخلات صور العمق الملونة بالإضافة لمستشعر لمسي في رؤوس الأصابع يفسر نقص المعلومات، حيث يرى أن الاستشعار اللمسي قد يوفر نهجا واعدا لمراعاة حالة عدم اليقين التي تنشأ عند محاولة روبوت إمساك جسم جديد تماما، ووجد أن هذه القضية لم تغطى بشكل واسع في الأدبيات إذ تفتقر الأيدي الروبوتية الحالية للتغطية اللمسية مما ينشئ مشكلة في قضية التخطيط لتنفيذ الإمساك فدرس كيف يمكن أن يساهم ذلك في أمثلة عملية التحكم. وصفت المشكلة المراد حلها بمرحلتين: 1. على الروبوت التخطيط لقبضة دقيقة تزيد فرص نجاح الإمساك للأجسام الجديدة. 2. تنفيذ هذه القبضة باستخدام تغذية عكسية لمسية تعوض عن حالة عدم اليقين المتعلقة بالشكل. ولتحقيق الهدفين اتبع نهج من أربع مراحل وهي: (1. التخطيط المسبق لوضعية الذراع واليد بتكوين مسبق مناسب للإمساك من خلال شبكتين تأخذان فوكسيل للكائن، 2. تنفيذ التشكيل المخطط له ثم التخطيط المسبق لتّحرك الأمثل للأصابع بحيث تتصل

مع الجسم في المناطق الأمثل لتحقيق قبضة ملائمة، 3. تنفيذ عملية الإمساك المخطط لها، 4. تقييم النماذج الخاصة بالتخطيط المسبق للتكوين وللإمساك بالتلامس مع مصنف القبضة الذي يتخذ قرار بشأن رفع الجسم أو إعادة الإمساك به). تستند المرحلتين (1) و(2) على مدخلات سحابة نقطية تقدر شكل الكائن، بينما تستخدم المرحلة (3) التغذية العكسية اللمسية لتنفيذ الخطة، وتستخدم المرحلة (4) المدخلات المرئية والتكوين الحالي للروبوت. أجريت التجارب الخاصة بالتدريب ضمن محاكي Gazebo باستخدام ذراع KuKa ذو 7 درجات حرية ويد Allergo ذات الأربع أصابع ولها 16 درجة حرية، استخدمت كاميرا لتوليد السحابة النقطية لكائن على الطاولة، ثم تم الاختبار ضمن المختبر بذات الإعدادات. أجري عددا من التجارب والتقييمات لمتحكمات التخطيط والتنفيذ، واعتبرت المهمة ناجحة فقط إذا استطاع الروبوت رفع الجسم لمسافة 15 سم تقريبا فوق الطاولة. كانت النتائج الكمية لنسبة نجاح النهج المقدم= 75% وبالمقارنة مع منهجيات سابقة: B_4 = 75% B_3 = 46% B_2 = 28% B_1 = 65% وكما هو واضح كان النهج المقدم أكثر دقة من منهجيات المقارنة، واستطاع العمل بشكل جيد في بيئات العالم الحقيقي رغم أنه درب فقط في المحاكاة. هذا البحث لم يستخدم التعلم المعزز العميق بل اعتمد على شبكات عصبية ومنهجيات حسابية أخرى مثل شبكة PointSDF للسحابة النقطية، وتنفيذ التخطيط الحركي باستخدام مخطط RRTConnect من MoveIt! لكنه أكد أن إضافة ردود الفعل اللمسية حسن من النتائج في الحصول على قبضات الإمساك المرغوبة بشكل ملحوظ (Matak et al, 2022).

2.3. دراسات تعتمد التغذية العكسية البصرية:

في بحث Odeyemi عام 2024 حول الإمساك الآلي بالأشياء للأيدي الاصطناعية الذكية استخدم ثلاث خوارزميات للتعلم المعزز وهي SAC و(Deep Q-Network - DQN) وPPO لتدريب الوكلاء على مهام الإمساك الآلية. دُرب مقبض بإصبعين متوازيين ضمن محاكي PyBullet حيث يقرأ الوكيل صور العمق الملونة (Red Green Blue depth - RGB-D) –، وينفذ إجراء على مفاصل الروبوت، أما تابع المكافأة فهو من نوع المكافآت المشكلة (Shaped Reward)، إذ تُمنح المكافآت أو العقوبات في كل حلقة خلال مسار التعلم اعتمادا على عوامل مختلفة، مثل ما إذا كان القابض قد أنجز مهمته أو المدة التي استغرقها لإنجاز تلك المهمة. أي يتم استخدام دالة المكافأة المشكلة للتكيف مع مهمة الإمساك المحددة بدلا من دالة المكافأة المتفرقة (Sparse Reward) التي تمنح مكافآت فقط عند إنهاء الحلقة. تنتهي الحلقة عندما يلتقط القابض الجسم بنجاح أو إذا لم يتمكن من ذلك بعد عدد محدد من الخطوات. بعد التدريب مع ألف حلقة لكل منها ألف خطوة بإجمالي مليون خطوة، تشير النتائج إلى أن خوارزمية SAC تحقق أعلى معدل نجاح بنسبة 99% بين الخوارزميات الثلاث عند أقل من 200 ألف خطوة زمنية في حين حققت DQN 61% وPPO حققت 83% (Odeyemi et al, 2024). يمكن أن يُعزى معدل النجاح المرتفع الذي حققته SAC إلى كونها OFF-Policy وإلى مزاياها المتأصلة في التعامل مع مساحات العمل عالية الأبعاد. يُظهر هذا

البحث أيضا أن الخصائص الفيزيائية للأشياء الملتقطة من خلال كاميرا العمق يمكن أن تؤثر على قدرة الوكيل على تعلم سياسة مثالية. كما يمكن أن يُعزى معدل النجاح المنخفض الذي حققته DQN إلى حقيقة مفادها أن بنية التعلم Q المستخدمة بواسطة خوارزمية DQN معروفة بامتلاكها مشاكل في الاستقرار عند تطبيقها على مهام تحكم مستمر. وعلاوة على ذلك، تسلط النتائج الضوء على إمكانات PPO في تطوير أيدي اصطناعية ذكية ذات قدرات إمساك تلقائية بالأشياء رغم أنها خوارزمية ON–Policy ما شجع على مزيد من البحث باستخدام هذه الخوارزمية.

في بحث Zheng وزملاؤه عام 2023 قُدم نهج لتعليم روبوت الإمساك باستخدام خوارزمية PPO مع $YOLOv_5$. أولا، يتم استخدام $YOLOv_5$ لتحديد الكائن والحصول على معلومات الموقع. ثانيا، يتم استخدام خوارزمية PPO للحصول على استراتيجية للإمساك. ثالثا، تتم مقارنة خوارزمية PPO مع خوارزميات SAC و (Trust Region Policy Optimization - TRPO) مع حجم دفعة 16 و128 على التوالي. تم العمل ضمن بيئة محاكاة MATLAB وباستخدام Python، كذلك طبقت التجارب على منصة فيزيائية لمركبة Mecanum تحمل ذراع DOFBOT يمكنها الإمساك بأشياء يصل وزنها إلى 500 جرام في دائرة نصف قطرها 30 سم. لتقييم النتائج تم تتبع المكافأة المتوسطة خلال تدريب الخوارزميات الثلاث. تظهر النتائج التجريبية أن الطريقة المقترحة، من حيث سرعة التعرف على الأشياء، تتفوق على نموذج $YOLOv_4$. يحقق نموذج $YOLOv_5$ دقة بنسبة 96% على مجموعة بيانات التعرف التي تم بنائها والتي تتمتع بدقة اكتشاف أعلى ومتطلبات أجهزة أقل من نموذج $YOLOv_4$. تفوق النهج المقترح على خوارزميات SAC وTRPO في الإمساك بالأشياء، حيث تم تحسين المكافأة المتوسطة لخوارزمية PPO بنسبة 93.3% و41% مقارنة بخوارزميات SAC وTRPO على التوالي. أخيرا، قارن الباحثون عملهم مع الأعمال ذات الصلة، ووجدوا أن النهج المقدم يتمتع بدقة أعلى ومتوسط دقة بقيمة 92.3%. في التجارب خارج المحاكاة وُجد أن معدل نجاح الإمساك يصل إلى 90% في بعض التجارب وحتى 100% في أغلب التجارب، مما يوفر استراتيجية بحثية جديدة للإمساك بالأشياء بواسطة يد روبوتية. وتظهر النتائج أن الطريقة المقترحة تتمتع بسرعة أكبر في التعرف على الأشياء ودقة اكتشاف أعلى (Zheng et al, 2023). أكد هذا البحث على فعالية PPO في مهام الإمساك الروبوتي ومن الضروري متابعة المقارنة مع خوارزميات أخرى.

أجرى Coskun دراسة عام 2022 بهدف تعليم ذراع آلي له خمس درجات حرية في نهايته قابض بإصبعين متوازيين الإمساك بجسم صلب ضمن محاكي Webots مع خوارزمية DDQN بحجم دفعة تدريب 16 ثم 32، ومشاهدات بصرية ملونة. نظمت المكافأة بحيث تكون عالية عندما تصل أصابع القابض إلى نطاقات المواضع المحددة، ويتلقى الوكيل مكافأة سالبة (عقوبة) عندما يكون خارج هذه النطاقات. يكون لكل إصبع إجراء منفصل. لوحظ أن حجم 16 مقارنة بـ 32 أعطى نتائج أفضل. وتظهر النتائج النوعية التي قدمتها الدراسة أن وكيل RL تمكن من التعلم بشكل جيد بما يكفي لإنجاز مهمة الإمساك (Coskun et al, 2022).

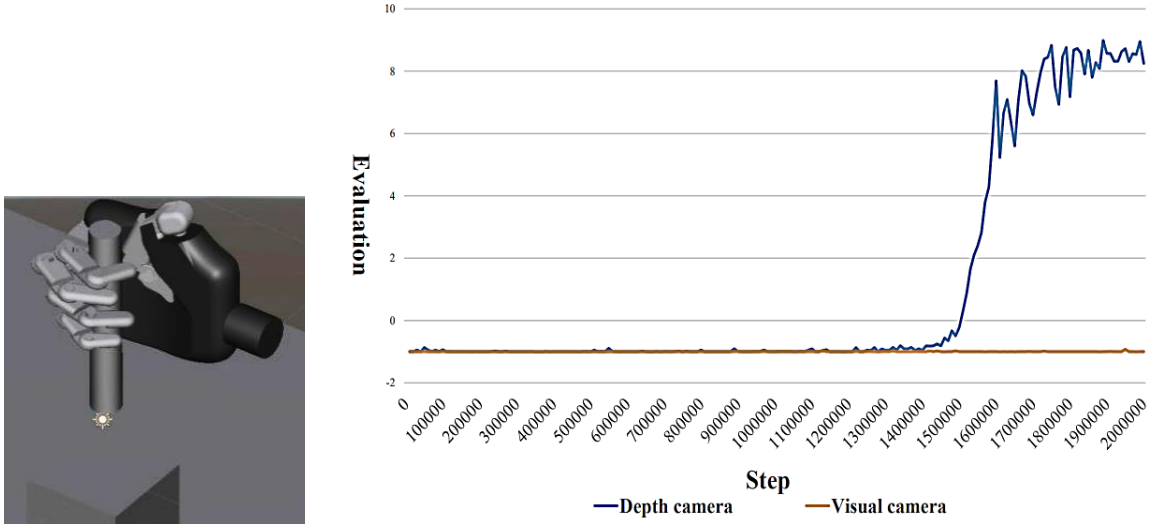
في بحث Imran عام 2020 استخدمت خوارزمية السياسة الحتمية العميقة (Deep Deterministic Policy Gradient – DDPG) مع صور العمق الملونة RGB-D لتدريب طرف اصطناعي علوي، تم استخدام محاكي MuJoCo مع اليد خماسية الأصابع SHAP والتي يتم التحكم فيها من خلال زوايا المحركات، يقرأ الوكيل معلومات من البيئة التي تضم حالة اليد والجسم المراد إمساكه، ويتم تنفيذ إجراءين: تحريك للقابض في الفضاء الديكارتي، أو أمر فتح وإغلاق لليد. يعطي هيكل مكافأة المتفرقة المستخدم 1 إذا نجحت اليد بالقيام بالمهمة و0 عكس ذلك. وتعتبر المهمة ناجحة عندما تستطيع اليد رفع الكائن فوق حد معين في نهاية الحلقة. استخدمت مجموعة بيانات صور عمق كبيرة لأجسام حقيقية مع روبوتات للتدريب، من أجل تمكن التعميم لاحقا للنموذج. بمجرد الحصول على إمساك ناجح للأسطوانة تم التعميم على أجسام جديدة لم يرها النموذج مسبقا، وكذلك بيئات فيها عدة أجسام. حقق النموذج معدل نجاح إمساك 78% (Imran et al, 2020). وقد يكون السبب في هذه النسبة المنخفضة استخدام البحث للمكافآت المتفرقة والتي من المعروف أنها تجعل من الصعب اكتشاف الحالات المرغوبة.

قام Chervinskii عام 2020 باستخدام التعلم المعزز العميق لتعليم روبوت (Universal Robots - UR5e) إدخال وتد في ثقب ضمن محاكاة CoppeliaSim ثم نُقل النموذج المتعلم إلى روبوت حقيقي دون أي ضبط إضافي، أظهرت السياسة المنقولة التي تأخذ فقط RGB-D أنها يمكن أن تعمل بشكل جيد على الروبوت الحقيقي. تتكون المكافأة الإجمالية لكل خطوة من المكونات التالية:

- مكافأة المسافة: بناءً على المسافة بين الوتد والهدف. وكلما كانت المسافة أصغر، زادت المكافأة.
- مكافأة النجاح: مكافأة متفرقة بناء على المسافة بين الوتد والموضع المطلوب في الفتحة. تتم مكافأة الوكيل بمجرد اقتراب الوتد كثيرا (حد معين يدويا) من الهدف.
- مكافأة الاصطدام: لتثبيط الروبوت عن الاصطدام بالكتلة، يحصل على عقوبة عند كل اصطدم.
- مكافأة الوقت: إجبار الوكيل على إكمال المهمة بأسرع وقت، يحصل على عقوبة صغيرة لكل خطوة.

تمت مقارنة 3 خوارزميات وهي: SAC وDDPG وخوارزمية السياسة المتدرجة الحتمية العميقة ذات البنية المزدوجة والمؤخرات الزمنية (Twin Delayed Deep Deterministic Policy Gradient - TD3) ولدت TD3 أكثر المكافآت وعند نقلها على الروبوت الحقيقي لم يتمكن نموذج صور العمق من تحقيق الدقة المرغوبة إلا أن استخدام الصور الملونة فقط أبدى أداء أفضل بكثير (Chervinskii et al, 2020). تبين نتائج الدراسة أنه يمكن تدريب وحدة تحكم تعتمد على العين في اليد فقط، في المحاكاة لإتمام المهمة بدقة عالية ونقلها إلى العالم الحقيقي. ومع ذلك، لا يزال هناك مشاكل عدم استقرار ناجمة عن تغيير ظروف الإضاءة وعوامل بيئية أخرى، والتي يمكن تحسينها من خلال تحسين العرض في المحاكاة، أو الاعتماد على مدخلات أخرى إضافية.

في دراسة Ayukawa عام 2021 بهدف تدريب يد روبوتية خماسية الأصابع بـ 6 درجات حرية (درجة حرية لكل إصبع وللإبهام حركة دوران جانبي إضافية) داخل بيئة Unity، على الإمساك ورفع أسطوانة بالاعتماد على مدخلات كاميرا العمق المثبتة في راحة اليد، وباستخدام خوارزمية PPO. يأخذ الوكيل مكافأة كلما ارتفع الجسم على المحور Y أكثر من 6 وحدات عن الأرضية ومكافأة 1− إذا كان الهدف أقل من 3. يأخذ مشاهدات صور العمق (84 ×84) بكسل ملونة، وزاوية الدوران للمفصل لكل إصبع، وتكون الأفعال ستة إجراءات ثني لكل إصبع وتدوير للإبهام وهي [3,3−] حيث لكل خطوة تُغير زاوية الإصبع 3°. تمت مقارنة مدخلات صور RGB التي لم تحقق أي نجاح مع صور RGB-D التي أظهرت نتائج عند مليوني خطوة كما في الشكل (4−2) (Ayukawa et al, 2021). كان لمجال التحكم المستمر أثر في نجاح هذا النموذج وتحقيق مكافأة تراكمية عالية وإتمام المهمة بنجاح إلا أن المدخلات البصرية وحدها تحد النتائج مقارنة بالحواس البشرية التي تعتمد بالدرجة الأولى على اللمس.



الشكل 4−2 نتائج التعلم النوعية والكمية ضمن المحاكاة (Ayukawa et al, 2021).

2.4. دراسات تعتمد أنواع أخرى من التغذية العكسية:

قام Gallouedec عام 2022 بتعليم روبوت إمساك الأجسام باستخدام DRL. استخدمت مجموعة من 4 بيئات محاكاة للروبوت ضمن بيئة تجارب OpenAI/gym المبنية على أساس MuJoCo، لتنفيذ أربع مهام وهي: وصول لنقطة وتحريك مكعب لوضع معين ودفع مكعب وإمساك مكعب ونقله لموضع معين مسبقا، لكل بيئة تابعي مكافأة واحدة متفرقة وأخرى كثيفة Dense Reward تأخذ المتفرقة 1 للإنجاز و0 عدا ذلك، تأخذ المكافأة الكثيفة مكافئة بين [1−0] اعتمادا على مدى قرب الوكيل من النجاح، عند الحصول على نتائج جيدة نُقلت المحاكاة لبيئة حقيقية وطبقت على روبوت Panda ذو 7 درجات حرية وقابضين متوازيين في نهايته يمكنه تطبيق قوة تصل حتى 70 نيوتن، استخدمت خوارزمية DDPG دون أي تعديل بالمحددات، بالنسبة للمهمة الأولى ساعدت المكافأة المستمرة بالتدرب بشكل أسرع، إذ يمكن بسهولة إثبات أنه كلما اقترب القابض من الهدف، يجب أن تكون المكافأة أعلى.

وبالتالي، فإنها توجه استكشاف النموذج نحو المكافأة العالية، وتظهر نتائج التجارب أن ثلاث ساعات من التدريب على حاسب تقليدي سمح لإنسان آلي مُحاكى بأداء مهمة استيعاب بمتوسط معدل نجاح 95%، بالنسبة للمهام التالية الأكثر تعقيد يصبح تحديد المكافأة أصعب وأكثر تأثيرا على النتائج. كما لم تتمكن الخوارزمية من حل المهام مع المكافآت المتفرقة. عند إضافة إعادة تشغيل الخبرة بأثر رجعي (Hindsight Experience Replay – HER) تمكن الوكيل بدأ من الحلقة الرابعة (Fourth epoch) تعلم مهمة الوصول مع مكافآت متفرقة والتي لم يتم تعلمها سابقا. ومع ذلك، فقد كان التعلم فاشلا في المهام الأخرى. لم يتم تعلم أي منهم، ولم يرتفع معدل النجاح على الإطلاق أثناء التدريب. تم إضافة التدريب المتوازي للتجارب مع وبدون HER لثامن وكلاء معا لوحظ أن الوكيل قد تعلم مهمتي الوصول والدفع جيدا مع وجود HER واحتاج مزيدا من التدريب لأداء مهمة إمساك ونقل المكعب. تظهر النتائج أن طريقة HER توفر حلا سريعا للاستغلال مقابل الاستكشاف لمقايضة البيئات ذات الأهداف المتعددة، فهي تسمح بالحصول على نتائج جيدة مع هذه المهام المعقدة كالإمساك والنقل (Gallouedec et al, 2022).

هدف Shahid وزملاؤه عام 2020 من خلال دراسته البحثية إلى تعليم روبوت Franka Emika Panda يُتحكم فيه بعزم الدوران وله سبع درجات حرية، بنهايته قابض بإصبعين متوازيين، تنفيذ الإمساك باستخدام التعلم المعزز المستمر من الصفر دون عروض توضيحية و/أو بيانات حقيقية ضمن محاكي MuJoCo، باستخدام خوارزمية PPO. تم التدريب على أكثر من 10 ملايين خطوة زمنية، بحيث تستمر كل حلقة لمدة 600 خطوة، أي 17 ألف حلقة تقريبا، فيمنح ذاك 2.5 ثانية لإنجاز المهمة، إذ تم تسريع المحاكاة بمعدل ألفي مرة أسرع من الوقت الحقيقي من أجل التدريب بشكل أسرع. وفي المهمة يُطلب من الروبوت الوصول إلى مكعب بحجم 6 سم والإمساك به ثم رفعه عن سطح التلامس. يأخذ الوكيل معلومات عن الروبوت والمكعب كالموضع وسرعات المفاصل وغيرها. وتتحكم الإجراءات الثمانية بسرعة وزاوية مفصل الإمساك. من أجل توجيه تعلم السياسة وتوفير ردود فعل متكررة للوكيل حول السلوكيات المناسبة استخدم تابعين مكافأة لتعلم مهمتين معا.

1. تابع المكافأة على مهمة الوصول (عند الخطوة t):

$$r_t = w_{dist}(r_{dist}) + w_{vel}(r_{vel}) + w_{grip}(r_{grip}) \qquad (2-1)$$

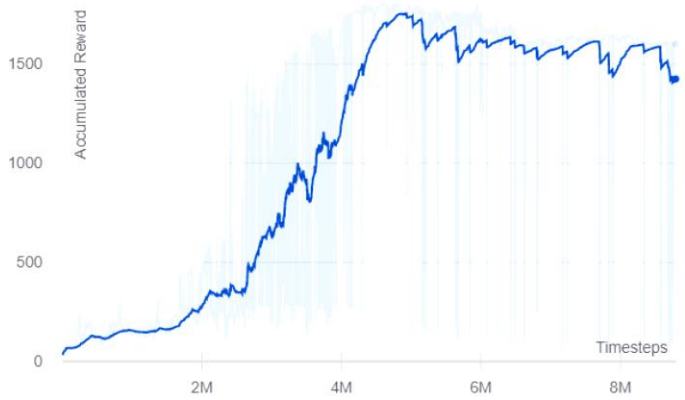
- r_{dist}: مكافأة المسافة المحسوبة بين المقبض والهدف.
- r_{vel} : مكافأة سرعة المقبض النهائي.
- r_{grip} : مكافأة فتح المقبض.
- أما w فهي توزين لكل مساهمة من المساهمات السابقة.

2. تابع مكافأة مهمة الرفع (عند الخطوة t):

$$r_t = (r_{dist}) + (r_{vel}) + (r_{grip}) + (r_{grasp}) + (r_{success}) \qquad (2\text{–}2)$$

- r_{dist}: مكافأة المسافة المحسوبة بين المقبض والهدف.
- r_{vel} : مكافأة سرعة المقبض النهائي.
- r_{grip} : مكافأة إغلاق المقبض.
- r_{grasp} : مكافأة إذا كان كلا اصبعي القابض على اتصال بالمكعب.
- $r_{success}$: مكافأة تعطى إذا تم الوصول للحالة النهائية المرغوبة.

تطلب التدريب حوالي 3.5 ساعة وكانت النتائج مشجعة إذ أن الوكيل الذكي قد تعلم كيفية أداء المهمة بنجاح بعد 4 ملايين خطوة كما يوضح الشكل (2–5)، لاحقا لذلك تم اختبار قدرات التعميم ومتانة سياسة الشبكة العصبية المكتسبة (النموذج المدرب) إذ تم إجراء 10 تجارب اختبار وتم تلخيص النتائج فيما يتعلق بالإنجاز الناجح للمهمة كما يلي، أظهر النموذج معدل نجاح 100% خلال 10 تجارب في التعامل مع عدة أشكال (مكعب 6 سم، مكعب 4 سم، أسطوانة بنصف قطر وارتفاع 3 سم، مفك براغي) موضوعة قريبا من الموضع الأصلي للمكعب، عند تغيير موضع الروبوت و/أو المكعب تناقصت معدلات نجاح الإمساك خلال 20 تجربة إلى قيم 7% – 8%. من أجل تكييف النموذج المكتسب مع تغييرات كبيرة في موضع المكعب تم إجراء تعليم إضافي إذ أضيفت 30 دقيقة من التدريب على النموذج المدرب، وعند الاختبار حقق الأداء نجاحا بنسبة 100%، مما يشير إلى أن السياسة يمكنها إعادة استخدام النموذج السابق بنجاح وتكييف السلوك بسرعة في المواقف الجديدة (Shahid et al, 2020). أبدت الدراسة نتائج مميزة بقدرة الخوارزمية PPO على تحقيق الإمساك مع روبوت له عدة مفاصل ومقبض ثنائي الأصابع لكنها لم تتعامل مع مشاهدات حسية أو حتى صور تقرب من الأداء الحيوي بل اقتصر الإدخال على بيانات من المحاكاة كموقع وسرعة المفاصل والجسم الهدف وهو ما يدفع لاستكشاف أدائها مع مدخلات أخرى، كما قدمت عدة مقاييس مهمة لنجاح الإمساك استفاد منها البحث الحالي.



الشكل 2–5 المتوسط المتحرك لقيم المكافأة التراكمية عبر الحلقات (Shahid et al, 2020).

في بحث Timmers عام 2018، استخدمت خوارزمية (Continuous Actor-Critic Learning Automaton – CACLA) لتعليم وكيل إمساك علبة مشروبات معدنية من خلال التحكم في سرعة محركات ذراع Mico ذات 6 درجات حرية مع قابض بأصبعين بدرجة حرية واحدة لا يحتوي أي مستشعرات ضمن محاكي V-REP. اعتمدت المكافأة على حسابات للمسافة بين القابض والكائن الممسوك وأضيف عقوبة على الاصطدام بالأرض. كان عدد الحلقات الأقصى 5 آلاف حلقة لكل منها 250 خطوة. بدأ التدريب أولا مع مفصل واحد لمهمة الوصول، ثم تمت إضافة مفاصل أكثر. لتعيين أفضل المحددات تجريبيا اختبرت معدلات تعلم مختلفة وأحجام طبقات متنوعة للشبكات. ثم تم الانتقال لمهام الإمساك دون الرفع، ويعتبر الوكيل ناجحا في الإمساك عندما تتلامس جميع الأصابع مع الكائن، ولتفادي مشاكل وقوع الجسم تم تبسيط الدراسة بفرض أن العلبة ثقيلة وساكنة ولا يمكن تحريكها، النتائج:

في تجارب مهمة الوصول: أثناء التدريب، تظهر منحنيات تتبع المتوسط المتحرك للمكافأة التراكمية مع مفصل واحد أن المكافأة التراكمية ترتفع بسرعة كبيرة وتبقى مستقرة، وعند إضافة المزيد من المفاصل يكون هناك المزيد من النتوءات للأسفل خاصة خلال أول ألف تجربة، يحدث هذا غالبا لأن الفاعل يبدأ بمخرجات سرعة منخفضة جدا ويتعلم ببطء التحرك نحو الهدف. ورغم أنه تعلم الذهاب إلى الهدف فإنه لم يتعلم بعد التوقف وبالتالي تجاوز الهدف وخلق مكافآت سلبية إضافية. أثناء الاختبار، تظهر النتائج على 10 تجارب أن معدل نجاح الانتقال مع مفصل واحد أعطى نسبة نجاح 100%. عند إضافة المزيد من المفاصل يمكن أن يتحرك المستجيب النهائي في ثلاث أبعاد ما يجعل حل المهمة مشكلة أكثر صعوبة، حيث أخذت معدلات النجاح تتناقص مع كل مفصل إضافي حتى وصلت إلى 0%. استُنتج مما سبق أن التعلم المباشر Online لا يكون فعال مع المزيد من المفاصل لذا تم تجربة وجود مخزن مؤقت يجمع تجارب على مدى 10 آلاف حلقة من المفاصل الستة، وتم اختبار هذه الطريقة مع 10 تجارب، حقق النموذج نسبة نجاح 90%. رغم النتائج الأفضل إلا أن النموذج استغرق وقت كبير تقريبا 10 ساعات للتدريب. تمكن الباحثين من خفضها لأربع ساعات مع استخدام التدريب المسبق المكون من 6000 تجربة يضاف لها 3000 تجربة تدريب، ورفع النتائج إلى 100%. لإظهار متانة الخوارزمية تم جعلها تنتقل لموضع عشوائي من أصل 100 موضع معد سابقا خلال 100 تجربة حققت نسبة نجاح 98% (Timmers, 2018).

في تجارب الإمساك: نظرا لأن الإمساك أصعب بكثير من مجرد الذهاب إلى نقطة ما، فتم زيادة إجمالي عدد الحلقات إلى 10 آلاف دون تدريب مسبق، حقق النموذج نسبة نجاح تصل إلى 70% يتضح أن الخوارزمية قادرة على تعلم الإمساك بالأشياء. الجانب السلبي هو أن التدريب يستغرق وقتا طويلا أكثر من 10 ساعات. لاحقا لذلك استخدمت شبكة عصبية لكل مفصلين (بالتالي 3 شبكات فاعل و3 شبكات ناقد) فتحسن معدل النجاح لـ100% مقارنة بشبكة فاعل – ناقدة واحدة لكل المفاصل، من سلبيات هذه الطريقة زيادة وقت التدريب 1.5 مرة تقريبا عن الشبكة الواحدة (Timmers, 2018).

قدمت الدراسة نتائج جيدة ولكن عدم الحصول على أي ردود فعل من الذراع لأنها غير نشطة جعل التعلم أكثر صعوبة، كما أنه يجب أن يؤخذ في الاعتبار أن الإمساك في الواقع لا يحقق الفرضية التي وضعها الباحثين بثبات العلبة الملتقطة ما سيسبب انخفاض بنقل النتائج لاحقا للواقع.

2.5. التعلم المعزز العميق والتغذية العكسية اللمسية في بيئات أكثر تعقيدا:

قام Puang وزملاؤه بدراسة عام 2024 بهدف تحقيق قبضات مستقرة للأجسام ذات مركز الثقل المتغير باستخدام معلومات لمسية فقط ونماذج المحولات (Transformers). طُلب من الوكيل التعلم من أجل تمكين قابض بدرجة حرية واحدة وإصبعين متوازيين إمساك ورفع حاوية مستطيلة الشكل فيها حمل متحرك على طولها. ويتطلب تحقيق هذا مراقبة سياسة التحكم لديناميكيات انتقال مركز الثقل والتنبؤ بموقع الإمساك الأفضل والذي يكون أقرب إلى المركز الجديد. كما أضيف بعدا آخر على استقرار القبضة بهدف التحكم بقوة القبضة الناتجة بحيث لا يتم تطبيق قوة مفرطة على الحمولة مختلفة الأوزان. نفذت التجارب ضمن بيئتين: المحاكاة الرقمية DiffRedMax، وأخرى واقعية (Puang et al, 2024). في المحاكاة، استخدم قابض بأصبعين متوازيين ومستشعر لمس على أحد الإصبعين لتقليل الحمل الحسابي وتم تعيين الوزن (25 أو 100) غرام، وعينت كثافته وحجمه وموقع البداية بشكل عشوائي. كما تم تعيين حد قوة الإمساك ليتطابق مع حد وزن الحمل، أما معامل الاحتكاك بين الحاوية والحمل يضبط عشوائيا بين قيمتين 0.11 و0.17 وذلك لتغطية شروط الانزلاق والثبات. تعتبر المهمة ناجحة إذا تمكن الروبوت من رفع الحاوية 20 مم لثانيتين ثم خفضها. نُمذج حساس اللمس بحجم 14×10 مم بخريطة لمسية بدقة 8×6 يحمل كل Taxel منها قيم قوى القص ثنائية البعد ويؤخذ 11 عينة منها أثناء الرفع لالتقاط الميلان وحركة الحمل، في مقياس الأداء يعتبر الإمساك ناجحا عندما يكون كل من الدوران والانزلاق للحاوية أقل من عتبتيهما، على الأكثر له انزلاق $\tau s = 3mm$ أي الرفع 17 مم على الأقل وإمالة $\tau r = 0.02\ rad$. وتقطع الحلقة وتعتبر فاشلة بعد 10 محاولات إعادة إمساك. أثناء التدريب استخدم محول ترميز Transformer Encoder لاستخراج السمات من الخرائط اللمسية، ومعالجتها وفهم الأنماط والعلاقات المعقدة بينها، ثم تستخدم خوارزمية SAC لتدريب سياسة الإمساك على تحدد موقع وقوة الإمساك. تم التدريب على التوازي لثمان بيئات لكل منها 90 ألف خطوة، استمر التدريب 2.5 ساعة. تُحدد المكافأة وفقا لحالتين: إذا كان الانزلاق أو الإمالة كبيرين سيحصل الروبوت على عقوبة سالبة تمنعه من فقدان الجسم. إذا كانت القيم ضمن الحدود المقبولة سيحصل الروبوت على مكافأة تعتمد على مدى مناسبة قوة الإمساك بالنسبة لوزن الجسم، مما يشجعه على استخدام القوة المثلى. فتقود الدالة سياسة التحكم إلى قبضة مستقرة عن طريق تقليل الإمالة والانزلاق للقضيب، وتحسن قوة الإمساك عن طريق تقليل القوة الزائدة δf. وبالتالي يكون هدفها: تجنب الانزلاق أو الإمالة الزائدة، واستخدام قوة إمساك مناسبة دون إفراط، مما يحسن كفاءة القبضة ويوفر الطاقة. تم تعيين معدل النجاح، ومتوسط عدد محاولات الإمساك كمقاييس أداء لمهمة الإمساك المستقرة، وأما لتقييم تحسن

قوة الإمساك حسب المتوسط الحسابي للقوة الزائدة المطبقة (القوة الزائدة / الخطوات). اختبرت السياسة في الواقع مع حلقة من 500 خطوة تظهر النتائج في الجدول (1–2) أن النموذج المقترح حل المهمة بمعدل نجاح يزيد عن 0.95 وبقوة زائدة أقل من 0.2 نيوتن دون الحاجة لتدريب إضافي. عند المقايضة بين هدفين: حد أدنى من محاولات الإمساك، وحد أدنى من القوة، وعلى الرغم من تحقيق النموذج المقترح للتوازن بين الهدفين وجد أن السياسة بشكل عام تميل للتركيز على هدف واحد أكثر من غيره حيث تنخفض نسبة النجاح (يزداد τ_s, τ_r) بتقليل الضغط وهذا منطقي، كما يميل النموذج القائم على المحول إلى الحصول على قوة زائدة أقل أكثر من تركيزه على عدد المحاولات. في حين عند مقارنة النتائج بين استخلاص الميزات القائم على المحول وبين استخلاص الميزات القائم على (Convolution Neural Network – CNN)، أعطى CNN معدل نجاح أكبر ويميل لأن يكون لديه عدد أقل من المحاولات في حين يميل النموذج القائم على المحول المقترح لأن يكون لديه قوة زائدة اقل. عند الاختبار في بيئة حقيقية دون إعادة تدريب نجح النموذج في الإمساك بالصندوق في جميع المحاولات الستين (Puang et al, 2024).

الجدول 1–2 مصفوفة الأداء (performance metrics)

Method	α (معلمة للمكافأة)	Success rate	Average #attempts	Average #excess force
SAC	10	0.990	3.071	0.271
	20	0.978	3.487	0.182
	30	0.954	4.172	0.137
CNN-based	10	0.992	2.887	0.821
	20	0.990	2.762	0.205
	30	0.986	3.302	0.194

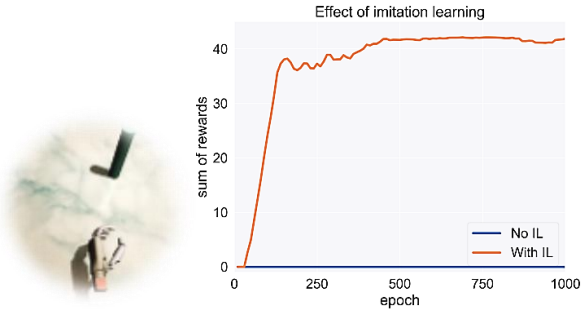
2.6. دراسات تعتمد على التعلم المعزز العميق مع العروض التوضيحية:

التعلم المعزز العكسي (Inverse Reinforcement Learning - IRL) هو أحد فروع RL حيث تُستنتج دالة المكافأة التي توجه سلوك الوكيل بالاستدلال عليها من عروض الخبراء التي تُعطى كعينات من السياسة المثلى أو شبه المثلى بدلا من تعريف المكافأة مسبقا وتجنب هندستها يدويا كما هو الحال في RL التقليدي. قد تمنح وظيفة المكافأة نفسها معلومات حول خصائص مهمة قد يتم تجاهلها في التصميم اليدوي. يُعرض في هذه الفقرة نتائج استخدام IRL في التدريب على الإمساك، ورغم أن نهج التعلم المحاكي من العروض التوضيحية أو ما يعرف بالتعلم العكسي، يبدي نتائج جيدة عادة عند التعامل مع درجات حرية عالية إلا أنه يعتمد على بيانات توضيحية تحتاج جهود كبيرة لجمعها وقدرات حسابية عالية، لذا سيتم دراسة تعلم مهام شبيهة دون عروض ثم تقارن النتائج.

درب Rivera وزملاؤه عام 2021 سياسة (Natural Policy Gradient – NPG) مع دمج فضاء التآزر الحركي Synergy Space كتمثيل منخفض الأبعاد لحركات اليد، بهدف إنشاء قبضات طبيعية، في كل خطوة يتم التحكم في زاوية المفاصل وموضع اليد وتقرأ السياسة هذه القيم مع المعلومات اللمسية يتم العمل ضمن بيئة محاكاة

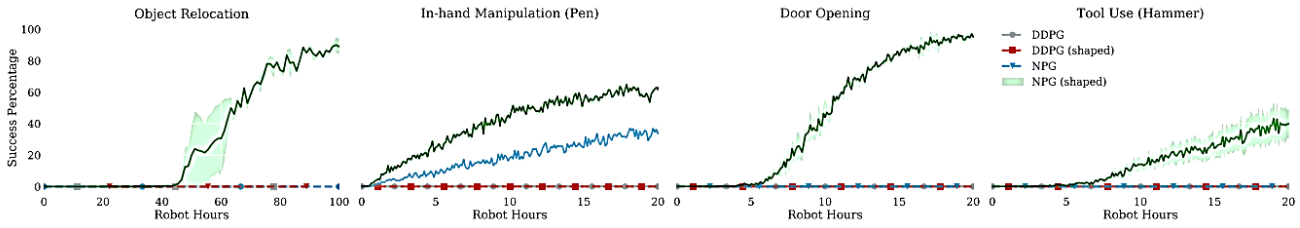
MuJoCo ويد ADROIT مع 22 درجة حرية، يحقق النهج المقترح القائم على التآزر معدل نجاح متوسط يبلغ 88.38% لمهام التحكم الدقيق بالأشياء، بينما يحقق النهج القياسي بدون مساحة التآزر 50.66% فقط. تُظهر النتائج النوعية أن السياسة المقترحة مع التآزر تنتج وضعيات أصابع تشبه حالة الإنسان على سطح عدة أجسام. وساعد التآزر في تقليل مساحة الإجراءات اللازمة لتعلم سياسة الإمساك. وإنتاج أوضاع يد وإمساك تشبه الإنسان (أي فتح اليد قبل الإمساك ووضع الأصابع على الكائن بطريقة طبيعية) (Rivera et al, 2021).

قدم Sharif وزملاؤه عام 2020 إطار قائم على RL والتعلم المحاكي (Imitation Learning – IL) لتوجيه استكشاف الروبوت في فضاء الحالات، وذلك بهدف التحكم في يد اصطناعية روبوتية لحالات بتر تحت الكوع مكونة من خمس أصابع لها خمس درجات حرية واحدة لكل أصبع، باستخدام مسار اليد البشرية فقط كمدخل. تم افتراض تآزر الأصابع بحيث تتحرك جميعا بالتساوي، يتم قياس مجموعة متغيرات وهي زوايا المفاصل وسرعتها ومسافات بين اليد والجسم. ويطلب رفع أسطوانة من خلال قبضة أسطوانية بشكل مستقيم عن حد معين. تكون الإجراءات سرعة تُقييس ضمن [1,−1] لجميع المحركات بالتالي فضاء الإجراءات هو فضاء متصل. يكافأ الإمساك الناجح بـ 1 لتشجيع الوكيل على الإمساك الجيد بأسرع ما يمكن و0 خلاف ذلك. تم استخدام خوارزمية SAC، ونظرا لصعوبة الاستكشاف استخدم IL لتوجيه النموذج واستكشاف الحالات المجزية بسهولة وبشكل أسرع. كما أضيفت بعض الخبرة من خلال متحكم برمجي سمي "الخبير" عبارة عن وحدة تحكم تناسبية تحصل على صفة اليد (إغلاق وفتح) وكذلك مسافة اليد النسبية عن الجسم كمدخلات وتعطي سرعات المفاصل كناتج، تحاول أولا فتح اليد ثم تبدأ الإغلاق عندما تكون المسافة أقل من عتبة محددة بحيث يمكن الإمساك بالكائن. تم إجراء التجارب في محاكي MuJoCo، كل حلقة تتألف من ألف خطوة وتم رسم المتوسط المتحرك للمكافآت بنافذة 15. تبين النتائج أن دمج RL + IL سمح بتعلم سياسة تكون دائما ناجحة ويحدث التقارب في الحلقة 500 تقريبا كما يوضح الشكل (2−6) تصل لقيمة 43 وهو أعلى مجموع ممكن من المكافآت الممكنة للحلقة الواحدة. بدون IL لا يتم تعلم أي سياسات مفيدة على الإطلاق (Sharif et al, 2020). يمكن أن يكون سبب عدم تعلم الوكيل لسياسة مجدية دون IL هو حقيقة أن الحالات المجزية يصعب استكشافها مع المكافآت المتناثرة في إعداد المشكلة ما يجعل طرق الاستكشاف العشوائية غير فعالة، كما أن المشاهدات مقتصرة على مجموعة قيم عديدة للزوايا والمسافات ما أثر على التعلم.



الشكل 2−6 الرسم البياني للتعلم لمشكلة RL الأصلية مع وبدون IL (Sharif et al, 2020).

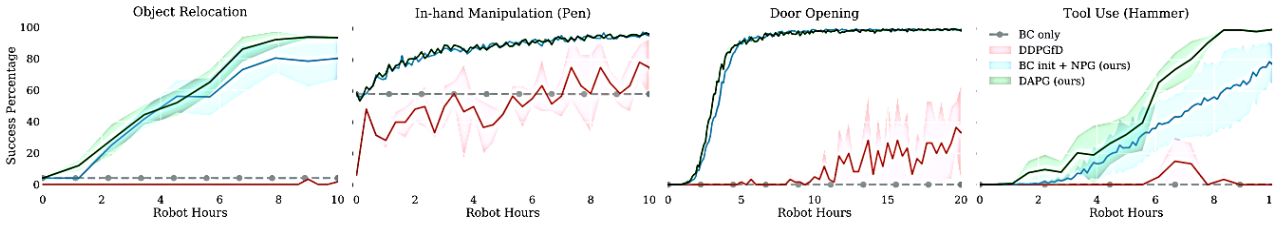
في دراسة Rajeswaran وزملاؤه عام 2018 بهدف تعليم اليد الروبوتية ADROIT خماسية الأصابع ذات 24 درجة حرية ضمن بيئة محاكاة MuJoCo أربعة مهام رئيسية باستخدام IRL، وهي مهام نقل الكائن، والتحكم اليدوي الدقيق، واستخدام الأدوات، والتحكم الدقيق بالأدوات البيئية الداعمة (كالأبواب). يتم التحكم بكل درجة حرية من خلال جهاز تحكم بالزاوية في المفصل وفيه مستشعر لقيمة الزاوية. جُمع 25 عرضا توضيحيا ناجحا، وزعوا عشوائيا لجميع المهام، كل عرض يتألف من مسارات الحالة اللازمة لأداء المهمة. تم تحديد زوايا مفاصل اليد وموضع واتجاه العنصر وموضع الهدف كمشاهدات من البيئة، وتغير الزاوية لمفاصل اليد كإجراءات. في التعلم من الصفر: تم تقييم خوارزمية NPG وDDPG مع مكافأة متفرقة، يبين الشكل (7–2) أن سياسات الاستكشاف العشوائي لا تحقق أي نجاح باستثناء مهمة الالتقاط باليد وبالتالي لا يتعلم الوكيل. عند تقييم التعلم من الصفر مع المكافآت المشكلة لدراسة مدى تأثيرها على النتيجة وتبين أن NPG قادرة على تحقيق نسبة نجاح عالية في هذه المهام بينما لم تكن DDPG قادرة رغم الضبط الكبير لمحدداتها، من المعروف أنها حساسة جدا للمحددات الفائقة والبذور العشوائية مما قد يفسر صعوبة تحجيمها إلى مهام معقدة وعالية الأبعاد. بالتالي تبين أن خوارزميات RL الحالية يمكنها حل المهام المحددة لكنها تتطلب جهد يدوي كبير في تشكل المكافأة، وتتطلب وقت يصل إلى 100 ساعة لبعض المهام، ونتج عنها استراتيجيات خاصة ذات صلابة ضعيفة (Rajeswaran et al, 2018).



الشكل 7–2 أداء خوارزميتي NPG وDDPG مع المكافآت المتناثرة والمُشكلة. إعداد المكافأة المتناثرة كان غير فعال معظم التجارب عدا تجربة التحكم اليدوي الدقيق، عند تضمين معرفة بشرية مسبقة بتشكيل مكافأة فساعد ذلك NPG على الانطلاق. بينما لاتزال DDPG تعاني للنجاح (Rajeswaran et al, 2018).

مع دمج العروض التوضيحية: في البداية استخدم التدريب باستنساخ السلوك (Behavioral Cloning – BC) فقط، والذي لم يبدي أي نتائج تعلم كما يوضح الشكل (8–2). ثم دربت سياسة NPG المهيئة مسبقا من خلال استخدام استنساخ السلوك (فيه تحاول السياسة تقليد الإجراءات المتخذة في العروض)، وفر هذا سياسة مستنيرة يتم توجيه استكشافها بكفاءة أكبر وقلل استخدام هذا النهج من تعقيد العينة لمهام التلاعب الماهر بشكل كبير، ولكن تبين أن السياسة المستنسخة لا تكن ناجحة عادة. ثم تم استخدام العروض التوضيحية لتهيئة NPG مع المكافآت المشكلة وسميت الخوارزمية (Demonstration Augmented Policy Gradient – DAPG). وجد أن خوارزمية DAPG تكتسب سياسات تعرض المزيد من الحركة الشبيهة بالإنسان، وأيضا أكثر قوة بكثير عند اختبارها على أشكال بخصائص فيزيائية (أحجام وكتل) مختلفة. كما وجد أنها يمكن أن تكوّن عينة أكثر كفاءة بما يصل إلى 30 ضعفا من RL من الصفر مع المكافآت المشكلة. وتكون قادرة على تدريب السياسات على المهام في أقل من 5

ساعات، والتي من المحتمل أن تكون عملية للتشغيل على أنظمة حقيقية. يتفوق DAPG على خوارزمية (Deep Deterministic Policy Gradient from Demonstrations – DDPGfD) في جميع المهام قبل أن يبدأ DDPGfD في إظهار علامات التقدم (Rajeswaran et al, 2018). رغم النتائج القوية التي قدمتها الدراسة إلا أنها لم تستخدم أي مدخلات بصرية أو لمسية لتدريب الخوارزميات والذي نعتقد أنه أمر مهم لأمثلة النتائج.



الشكل 8–2 أداء خوارزميات RL مع العروض التوضيحية – DAPG (الخاصة بالدراسة) وDDPGfD. يحقق DAPG أداء أفضل مقارنة بـ DDPGfD. بالنسبة لـ DAPG رُسم أداء السياسة العشوائية المستخدمة في الاستكشاف وفي أي حلقة يكون أداء السياسة الحتمية الأساسية أفضل (Rajeswaran et al, 2018).

يستعرض الجدول (2-2) عدد من الدراسات الأخرى التي لا مجال للتوسع في ذكرها في هذه الرسالة، والتي تبين الاتجاهات الحديثة في استخدام خوارزميات التعلم المعزز العميق والتعلم المعزز المحاكي لتعليم الأيدي الروبوتية مهام الإمساك تحت ظروف مختلفة من التحديات.

الجدول 2–2 ملخص لبعض الدراسات السابقة الأخرى.

العام – المرجع	المهمة	المنهجية	نسبة النجاح
(Zeng et al, 2018)	الدفع والإمساك	درب روبوت UR5 مع قابض ثنائي الأصابع في بيئة محاكاة V–REP وفي الواقع مع مشاهدات بصرية، باستخدام خوارزمية Q – Learning.	65.4%
(Joshi et al, 2020)	الإمساك	استخدم المشاهدات البصرية للتحكم في روبوت Baxter بحيث يتمكن من الإمساك بمجسمات بسيطة ضمن بيئة Gazebo وعلى الروبوت الحقيقي، باستخدام خوارزمية DDQL مع Grasp Q Network مقترح ضمن الدراسة.	87%
(Hundt et al, 2020)	الدفع والإمساك	اقترحت الدراسة استخدام إطار عمل جدولة المهام الإيجابية – (Schedule for Positive Task SPOT) مع خوارزمية Q–Learning حيث استخدمت المشاهدات البصرية للتحكم في قابض ذو أصبعين لأداء مجموعة متنوعة من المهام، ضمن بيئتي محاكاة وحقيقية.	76%
(Mohammed, 2020)	الإمساك والتموضع	درب روبوت UR5 مع قابض ثنائي الأصابع في بيئة محاكاة V–REP مع مشاهدات بصرية، باستخدام خوارزمية Q –Learning.	74.3%
(Al–Shanoon et al, 2021)	الإمساك	استخدام المشاهدات البصرية للإمساك بأجسام بسيطة بدأ من الصفر لروبوت UR5 ذو 6DOF، مع مقبض بأصبعين باستخدام (Deep Reinforcement Grasp Policy – DRGP) ضمن بيئة محاكاة V–REP وفي تجارب على الروبوت الحقيقي.	62%
(Peng et al, 2022)	الدفع والإمساك	استخدمت DQN مع كاميرا عمق ضمن V–REP لتدريب UR5 وقابض ثنائي.	83.5%
(Wu et al, 2023)	الدفع والإمساك	درب UR5 مع قابض ثنائي الأصابع ضمن V–REP وUR3 مع robotic–85 في الواقع مع مشاهدات بصرية، باستخدام DQN مما يساعد في أمثلة اتخاذ القرار.	75.3%
(Wang et al, 2024)	التكديس والإمساك	استخدم Q –Learning لتعلم ثلاث مهام وهي الإمساك، والتكديس، والدفع، لروبوت UR5 مع قابض ثنائي الأصابع ضمن V–REP مع مشاهدات بصرية.	86%

(Mohammed, 2021)	الدفع والإمساك	لتحقيق المهمة في سيناريوهات الفوضى والمرتبة والانسداد. قسم النهج إلى: (1) استخدام كاميرات متعددة لإعداد وجهات نظر متعددة لمعالجة مشكلة الانسداد؛ و(2) استخدام مشاهدات التغيير البصري على أساس اختلاف عمق البكسل. ضمن V-REP باستخدام UR5 مزود بمقبض فك متوازي.	**83.6%الازدحام 86.3%المنظمة 100%الانسداد**
(Franceschetti et al, 2020)	الإمساك والرفع	استخدم UR5 مع قابض بثلاث أصابع في المحاكاة والواقع. ضمن محاكي MuJoCo مع TRPO وDQNNAF وDDPG وVPG. لإمساك أسطوانة طولها 2 سم وكتلتها 0.5 كغ ثم وضعها على ارتفاع 30 سم فوق الجسم. في كل حلقة يُأخذ موضع المفاصل. أظهرت VGP وDDPG وDQNNAF تعلم غير مستقر وفشل كبير بسبب انزلاق الأسطوانة أو سرعة الاقتراب. بينما أظهرت TRPO بعد 5000 حلقة إمساك ممتاز. ثم نجح نشر السياسة دون الحاجة لإجراء أي تغيرات.	**100%**
(Nia et al, 2022)	الإمساك	تدريب نموذج يد عضلي هيكلي يحتوي على 29 عظمة و23 مفصلا و39 وحدة عضلية للتحكم في الوصول إلى وضع ثابت والإمساك بكرة في البيئة الافتراضية. تم تطبيق خوارزمية SAC في بيئة افتراضية من MyoSuite. تم تقييم التعلم تقييم نوعي وكمي لقيم المكافأة التراكمية.	**لم تذكر نسبة محددة ولكن تم التأكيد على نجاح المهمة.**
(Korkmaz et al,2018)	الإمساك	تعليم يد روبوتية ثلاثية الأصابع بتسع درجات حرية إمساك كرة كريكت على الطاولة المقابلة باستخدام التعلم المعزز المنفصل للتحكم في زوايا المفاصل ضمن محاكي Gazebo. استخدم روبوت UR-10 مع يد Shadow، وتم تطبيق خوارزميتي DQL و(Advantage Actor – Critic – A2C) على مفاصل اليد بالترافق مع استراتيجية الجشع المزدوج Double ϵ- Greedy. يقرأ النموذج مفاصل اليد ويقوم بإجراء تعديل على الزوايا وتحسب المكافأة بناء على جودة الإمساك مع وجود عقوبة في حالات الفشل. تظهر النتائج أن A2C تفوقت من حيث عدد الخطوات اللازمة لتحقيق النجاح (61 خطوة مقابل 113.4 لـDQL). كما أوضحت التجربة أهمية طريقة الجشع المزدوج في تحسين الاستقرار، حيث فشل النموذج عند غيابها. على الرغم من نجاح اليد الثلاثية في تنفيذ المهمة، إلا أن الدراسة لم تعتمد على أي مدخلات حسية من البيئة، واقتصرت على تتبع ارتفاع الكرة فقط.	**تعلم المهمة بنجاح لكن لم يذكر نسبة محددة لنتائج الاختبار**

2.7. خاتمة:

وأخيرا، شكّلت الدراسات التي تم تناولها في هذا البحث الأساس المتين الذي بُنيت عليه هذه الدراسة، حيث وفرت الإطار النظري والمنهجي اللازم لفهم الموضوع بعمق، وساهمت بمعرفة آخر التطورات في هذا المجال. كما وتعزز هذه المراجعة العلمية أهمية البحث الحالي وساهمت في توجيه مساره نحو تحقيق نتائج أكثر دقة وموثوقية.

3. الفصل الثالث: اليد البشرية والأطراف الاصطناعية: (دراسة بيولوجية ووظيفية لآليات التحكم والإمساك).

3.1. تمهيد:

بهدف تطبيق التعلم المعزز في الأطراف الاصطناعية لا بد من مراجعة وتذكر أهم الأسس النظرية اللازمة لتنفيذ البحث، في هذا الفصل توصف الخلفية النظرية الضرورية لفهم كيفية التحكم في اليد البشرية والأطراف الاصطناعية التعويضية الحالية، ويعتبر ذلك ضروريا لأن خوارزميات DRL تقلد النموذج الحيوي بالتالي لابد من فهمه لفهم مبدأ DRL، كما أن مراجعة طرق التحكم المتاحة حاليا ستوضح أهمية البحث الحالي بشكل أفضل.

3.2. يد الإنسان (دراسة بيولوجية):

اليد هي أهم العناصر التفاعلية لدى البشر، في الحياة اليومية وأثناء أداء المهام والحركات الدقيقة باستخدام الأيدي فإن مناطق مختلفة من الدماغ كمناطق الإدراك والمعرفة والتحكم الحركي بالإضافة إلى وحدات الاستشعار ومجموعة من الهياكل الجسدية تتعاون بشكل محكم ودقيق لتوليد القبضات المرغوبة. وقبل تصميم أطراف اصطناعية تحاكي هذا التحكم لابد من مراجعة الدراسة التشريحية والفيسيولوجية والسلوكية المتعلقة بالتحكم اليدوي الدقيق لدى البشر.

3.2.1. نظام الإدراك والمعرفة في اليد:

تدرك اليد البشرية المعلومات البيئية من المناطق المحيطة بها من خلال ثلاث جوانب وهي:

1) المستقبلات الميكانيكية والحرارية الموجودة في الجلد.
2) المدخلات الحسية العميقة الموجودة في العضلات والمفاصل.
3) الإشارات الصادرة من المراكز العصبية.

ولأداء مهام وحركات يدوية دقيقة ومعقدة قد يُجمع بين هذا الإدراك والبصر، حيث يوفر الحس البصري معلومات عن الشكل والفئة والتموضع والسرعة والقدرة على التحمل وما إلى ذلك بهدف التخطيط للمهمة وأداء الحركات الماهرة المطلوبة. بينما يوفر الحس اللمسي اتصالاً ديناميكياً وردود فعل عن القوة في الوقت الحقيقي أثناء عملية الإمساك وهو مفتاح أساسي لعمليات التحكم الدقيق والمرن والقوي والمتقن بالكائنات. وتصبح المستشعرات اللمسية أكثر أهمية عند غياب المعرفة العامة بالبيئة أو غياب الحس البصري كما في البيئات المجهولة والمظلمة (Li et al, 2022).

3.2.1.1. الاستشعار اللمسي في اليد:

تعتبر الأيدي نظام حسي معقد وحساس للغاية، وهي أكثر تعقيدا من الأذرع بسبب بنيتها التشريحية وإمداداتها العصبية. يتم تحقيق الإحساس اللمسي كجزء من نظام التغذية الراجعة اللمسية في الأيدي من خلال مستقبلات حسية متخصصة تقع تحت الجلد. مكّنت سنوات من البحث من التعرف على البنية التشريحية لجلد اليد والذي يتكون من ثلاث طبقات: البشرة والأدمة والطبقة تحت الجلد بدأ من الطبقة الخارجية إلى الداخلية (Li et al, 2022). أسفل طبقات جلد اليدين توجد أربعة أنواع من المستقبلات اللمسية تتوزع على النحو التالي:

1) خلايا ميركل (Slowly Adapting -SA1): تستخدم للإحساس بالملمس والضغط، وتحتوي على حقول استقبال صغيرة (Receptive Fields - RFs). وتستجيب لتشوهات وضغوط الجلد الثابتة بشكل مستمر.
2) جسيمات روفيني (SA2): حساسة للتشوه المستمر في الجلد وتنتج استجابة مستمرة أيضا، وتستخدم للكشف عن تغيرات الضغط المستمر والانزلاق.
3) جسيمات مايسنر (Rapidly Adapting -RA1): تحتوي على حقول استقبال صغيرة وتتحسس التحولات أو الاهتزازات السريعة وتنتج استجابات طوريه لتشوه الجلد، وهي سريعة التكيف.
4) جسيمات بانكيني (RA2): تملك حقول استقبال كبيرة وتكون حساسة للاهتزازات عالية التردد وتستخدم بشكل رئيسي في التفاعل مع الأدوات أو الكشف عن الاهتزازات السريعة (Li et al, 2022).

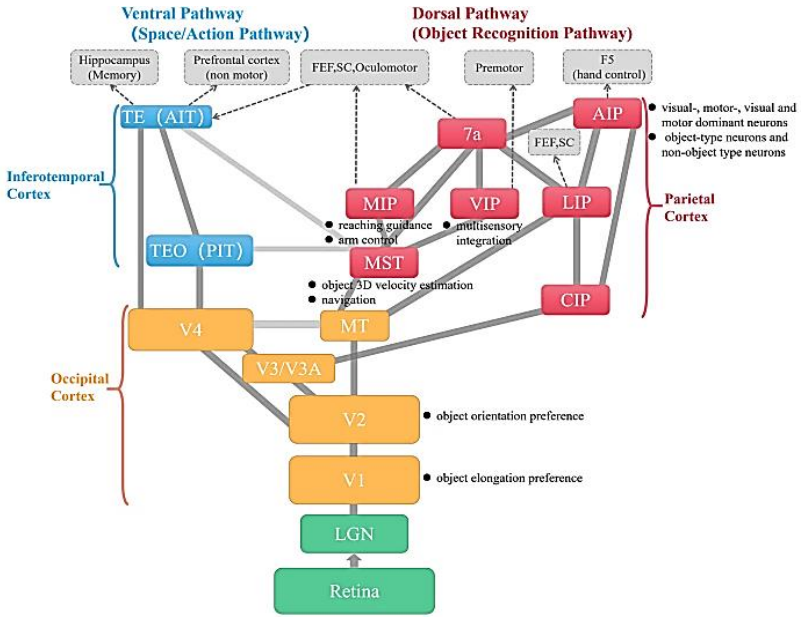
في مهارات التلاعب اليدوي تستخدم SA1 بشكل كبير كمدرك لحصول اللمس، في حين تستخدم SA2 للكشف عن الانزلاق، وتستخدم RA1 للكشف عن الحركة، وتساعد RA2 بشكل كبير في عملية استخدام الأدوات. إن الضفائر الليفية التي تنقل الإشارات الكهربائية العصبية الحركية أو الحسية تكون مدمجة داخل الطبقة تحت الجلد. وتشكل هذه الألياف شبكة كبيرة ومعقدة تربط المستقبلات الحسية أو النهايات العصبية الحرة بالحبل الشوكي أو الجهاز العصبي المركزي. وبالتالي، هناك عدد كبير من المستقبلات الميكانيكية المتصلة بالضفائر الليفية. يُظهر تقرير أن هناك ما يصل إلى 500 مستقبل ميكانيكي لكل سنتيمتر مكعب من اليد البشرية في المتوسط، وبالتالي تحقيق رد فعل لمس حساس للغاية مع دقة لانزياح الجلد تصل إلى 10 نانومتر تسمح للدماغ بمعالجة الإشارات الحسية بسرعة وبدقة، ما يجعل اليد قادرة على استشعار التغيرات والحركات الدقيقة جدا في الجلد (Li et al, 2022).

3.2.1.2. الاستشعار البصري وآلية معالجة الدماغ لبياناته دعما للتفاعل اليدوي مع الأشياء:

تملك القشرة البصرية في الدماغ مسارين بصريين لمعالجة الرؤية وهما المسار البطني (ventral pathway) والمسار الظهري (dorsal pathway)، يوضح الشكل (1−3) هذه المسارات. يكون المسار البطني مسؤولا عن التعرف على الأشياء وتمييزها إذ يساعد بمعرفة شكل ولون وخصائص الكائن ويعرف بمسار "ماذا". في حين يملك المسار الظهري مسؤولية تحديد موقع الأشياء في البيئة المحيطة ويقود حركة اليد للوصول والإمساك ويعرف بمسار "أين". بالنسبة للحركات اليدوية المعقدة المرتكزة على الأشياء كالإمساك الماهر والمتقن باليد يمكن أن يؤدي التفاعل بين المسارين إلى تعزيز وظائفهما بشكل متكامل حيث يوفر المسار الظهري توجيها بصريا لأنشطة التفاعل بين الإنسان والكائن كالوصول والإمساك ويحتاج معلومات تفصيلية وإرشادات بصرية وخصائص تفصيلية عن هوية الكائن تُسترجع من المسار البطني لكي يضبط القبضة بدقة، كما يأخذ المسار البطني من المسار الظهري تحديثات الحالات المتعلقة بالقبضة لتحسين التمثيل الداخلي للكائن ضمنه (Li et al, 2022).

3.2.1.3. تعديل الرؤية المرتكز على اليد:

تشير الدراسات العصبية الفسيولوجية والسلوكية إلى أن الدماغ يعدل طريقة معالجة الصور البصرية عندما تكون اليد قريبة من الجسم أو تتعامل مع الأشياء يساعد هذا التعديل في اختيار أفضل الخصائص للأشياء كـ (الحجم والشكل والاتجاه) بدقة أكبر أثناء التفاعل. مثلا تقوم منطقة القشرة البصرية الثانوية (Secondary Visual Cortex - V2) الموضحة في الشكل (1–3) وهي جزء من قشرة الدماغ البصرية بالتعامل مع المعلومات الخاصة بتحديد الاتجاه المفضل للأشياء أي كيف يجب أن تكون موضوعة لتكون قابلة للإمساك بشكل أفضل، وتأخذ هذه المنطقة المعلومات من منطقة القشرة البصرية الأولية (Primary Visual Cortex – V1) وهي أول منطقة في معالجة المعلومات البصرية وتقوم مع منطقة القشرة الجدارية–القذالية العلوية الخلفية (posterior Superior Parieto-Occipital Cortex - pSPOC) باختيار اتجاه المعصم، وتستخلص القشرة الجدارية–القذالية العلوية الأمامية (anterior Superior Parieto–Occipital Cortex – aSPOC) الخصائص البصرية للأشياء، ويُرمّز شكل الأشياء وعدد الأصابع في الثلم داخل الجداري الجانبي الأمامي (anterior lateral Intraparietal Sulcus – alPS). وتكون المنطقة الأمامية من الثلم الجداري الداخلي (Anterior Intraparietal – AIP) مسؤولة عن التعرف على الأشياء ومراقبة قبضة اليد. في الوقت نفسه تعد المنطقة البصرية 6A (Visual area 6A - V6A) كجزء من المسار الظهري الأنسي من القشرة الجدارية الخلفية لها وظيفة مشابهة لوظيفة AIP ولها دور حاسم في تكامل الرؤية مع حركة اليد. قد يكون الفرق بينهما هو تقسيم العمل، حيث أن AIP مسؤول عن التعرف على الأشياء والمراقبة البصرية للإمساك، بينما V6A تشارك بشكل رئيسي في التوجيه البصري لليد للوصول والإمساك. هناك فرق آخر هو أن AIP أكثر ارتباطا بالتحكم البطيء والدقيق بينما V6A تشارك في التحكم السريع والواسع (Li et al, 2022).



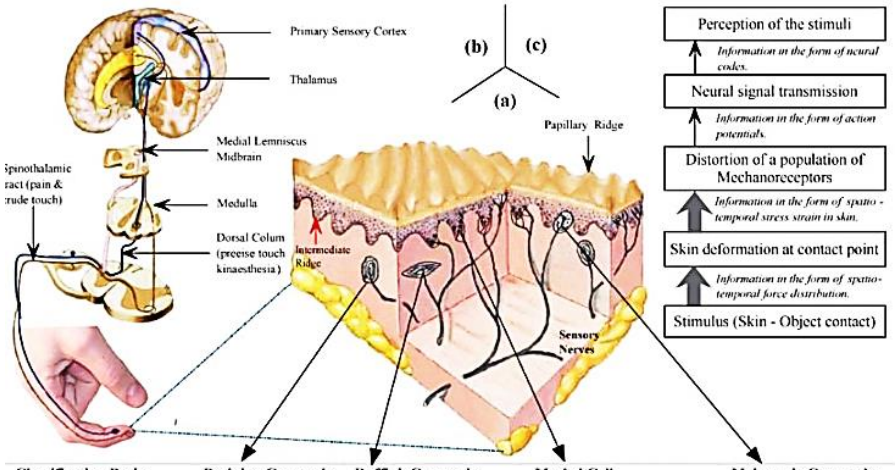
الشكل 1–3 المسارات البطنية والظهرية في قشرة الدماغ البصرية والمناطق المتصلة بها، باستثناء الشبكية ونواة الجسم اللاتيني باللون الأخضر، تمثل الكتل الصفراء والحمراء والزرقاء مناطق القشرة القفوية والجدارية والصدغية السفلية على التوالي. الكتل الرمادية مناطق أخرى من قشرة الدماغ تتعلق بالإدراك الحركي والتعلم والتحكم. تشرح النصوص السوداء وظيفة كل كتلة في تحقيق التلاعب (Li et al, 2022).

3.2.1.4. آلية معالجة اللمس في الجهاز العصبي البشري (مسار إشارات اللمس والوظيفة):

عندما نلمس شيء، تقوم المستقبلات الحسية (أو الميكانيكية) وشبكات الألياف العصبية (الضفائر الليفية) في الجلد بنقل الإشارات إلى الحبل الشوكي أو جذع الدماغ عبر الخلايا العصبية المستقبلة، تحتوي هذه الخلايا على:

1) جسم الخلية العصبية الواقع في العقدة الجذرية الظهرية للفتحة المخروطية للنخاع الشوكي.
2) فرع مركزي يتحد مع الجذر الظهري ويتجه نحو الحبل الشوكي.
3) فرع محيطي يندمج مع ألياف عصبية أخرى لتكوين عصب محيطي وينتهي بمستقبل متخصص في الجلد.

عادة ما تكون المحاور المحيطية طويلة وتولد جهود فعل (إشارات كهربائية) قريبة جدا من نهاية العصب وتنتقل عبر سيتوبلازما الخلية العقدية إلى الحبل الشوكي أو جذع الدماغ. تكشف تسجيلات الأعصاب الدقيقة أن معظم الألياف العصبية ليس لديها نشاط تلقائي كبير بل تُفعل فقط عندما يتم تحفيز الجلد. فعندما تلمس اليد شيئا يتم ضغط الخلايا داخل المستقبلات الحسية وتتحرك الطبقات الخلوية مما يولد إشارات كهربائية عصبية تنتقل عبر شبكة الألياف إلى الجهاز العصبي المركزي. بالإضافة إلى ذلك، في حالة التعامل اليدوي تكون حقول الاستقبال اللمسية في القشرة الحسية الجسدية (Somatosensory Cortex) أصغر حجماً من نظيرتها في القشرة الحركية والمناطق الخاصة بالإحساس الذاتي (الحس العميق). وبما أن المستشعرات الجلدية ترمز خصائص الشكل المحلي مثل التقوس واتجاه الحافة عند نقطة التلامس لذا فإن صغر حقول الاستقبال تسمح بتمثيل أكثر دقة للمعلومات اللمسية المكانية. بينما حقول الاستقبال في القشرة المحركة ومناطق الإحساس الذاتي تشمل عدة مفاصل تمتد على كامل اليد لذا تكون أقل دقة لكنها تفيد في الحركة العامة لليد. يوضح الشكل (2–3) هذه الآلية (Li et al, 2022).



Classification Basis		Pacinian Corpuscle	Ruffini Corpuscle	Merkel Cells	Meissner's Corpuscle
Type		FA II	SA II	SA I	FA I
Adaptation Rate		Fast	Slow	Slow	Fast
Spatial Acuity (mm)		10+	7+	0.5	3-4
Vibration/rapid indent threshold	Best(μm)	0.01	40	8	2
	Mean(μm)	0.08	300	30	6
Stimuli Frequency (Hz)		40-500+	100-500+	0.4-3	3-40
Conduction Velocity (m/s)		35-70	35-70	40-65	35-70
Effective Stimuli		Temporal changes in the skin deformation	Sustained downward Pressure; Lateral skin stretch; Skin slip	Spatial deformation; Sustained pressure; Curvature, edge, corners	Temporal changes in skin deformation
Sensory Function		High frequency vibration detection; Tool use.	Finger position; Stable grasp; Tangential Force; Motion direction	Pattern/form detection, texture perception; Tactile flow perception.	Low frequency vibration & motion detection, Grip control; Tactile flow perception

الشكل 2–3 توضيح لعملية نقل الإشارات اللمسية (James, 2021).

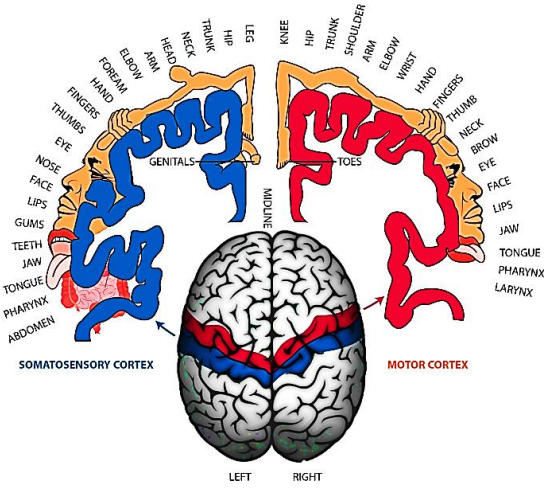
3.2.1.5. الاندماج البصري اللمسي في الدماغ أثناء التحكم باليد:

يحتوي دماغ الإنسان على خلايا عصبية يتم تنشيطها بمحفزات بصرية ولمسية وجسدية مثل المنطقة البطنية للثلم الجداري الداخلي (Ventral Intraparietal area – VIP) وقشرة الدماغ قبل الحركية (premotor cortex – PMC). ويتم دمج المدخلات متعددة الحواس السابقة بطرق خطية أو لا خطية. يساعد الدمج اللا خطي على تعزيز الإدراك العام وأداء الحركة بمزيد من الكفاءة والدقة. مع هذا الدمج يمكن للدماغ اكتشاف الإشارات الضعيفة أو التي تقع تحت العتبة، كما يمكن تعزيز بروز إشارة في إحدى الحواس بواسطة حاسة أخرى من خلال تعديل الانتباه. يحدث التعزيز وتظهر فعالية الدمج فقط عندما تكون الإشارات متزامنة مكانيا وزمنيا بالنسبة للحدث المستهدف (مثلا رؤية ولمس ذات الكائن بذات اللحظة) ويفسر ذلك بتداخل المحفزات متعددة الحواس داخل نفس حقل الاستقبال العصبي. أما إذا لم تكن الإشارات متزامنة فإن تنشيط الخلايا العصبية قد يكون أضعف حتى من حالة التحفيز الفردي أي قد يقل تنشيط الدماغ عوضا عن التحفيز. في عمليات الوصول والقبض باليد يتم تعديل الإدراك متعدد الحواس هذا بشكل ديناميكي حيث يتم تثبيط الإحساس اللمسي بقوة أثناء حركة الوصول باليد لتسهيل معالجة إشارات الإحساس الجسدي ثم يعاد التنشيط عندما تشارك الأصابع في عملية القبض. كما يمكن أن تساعد المعلومات البصرية الناتجة عن رؤية اليد في عملية الدمج البصري اللمسي مما يجعل من الصعب تحديد أي حاسة تعمل بالضبط في وقت معين نتيجة للتداخل الكبير. وباستثناء التحول في الحاسة المسيطرة عندما تكون الحواس الأخرى غير متاحة حاول بعض الباحثين فهم كيفية دمج الحواس المتعددة عندما تكون هناك معلومات متكررة عن ذات الكائن من أكثر من حاسة كالمعلومات البصرية واللمسية مثل الموقع والحجم وقد وجدوا أنه عندما تكون المعلومات عن الموقع والحجم موجودة من حاستي الرؤية واللمس فإن الدماغ يدمج فقط المعلومات اللمسية المرتبطة بالموقع مع المعلومات البصرية، ولا يدمج المعلومات اللمسية الخاصة بالحجم مع المعلومات البصرية مما يؤدي إلى حركات أسرع وفتح أقل لقبضة اليد، وتقدير الحجم بصريا بشكل أدق. علاوة على ذلك، درس بعض الباحثين الاختلافات بين اليد اليمنى واليسرى، ووجدوا أن اليمنى تفضل القبض الموجه بصريا للحركات الدقيقة، بينما اليسرى تتخصص في التعرف على الأشياء وتحديد موقعها بواسطة اللمس. من ناحية علم السلوك الحيواني والإحصاء وجد أنه بمجرد مراقبة حركة يد ما دون اللمس الفعلي يمكن للبشر أو الخوارزميات تقدير الخواص اللمسية للأشياء (Li et al, 2022).

3.2.2. النظام الحركي لليد:

يتم تنظيم التحكم الحركي ليد الثدييات في وضع هرمي، مما يعني أن الحركات يتم تنسيقها من خلال عدة مستويات بدأ من المستوى الأساسي وصولا إلى المستوى الأكثر تعقيدا. وفي خريطة (homunculus) التخيلية الموضحة في الشكل (3–3) تُمثل اليد كمؤثر في كل من القشرة الحركية والحسية الجسدية وتحتل مساحة كبيرة نسبيا في القشرة

الدماغية رغم صغر حجمها على مستوى الجسم، مما يعكس أهميتها البالغة ومستوى التحكم والإحساس العالي فيها. تتناول هذه الفقرة آليات التحكم الحركي والمكونات التي تساهم في تخطيط حركة اليد وتنسيق حركة الأصابع.



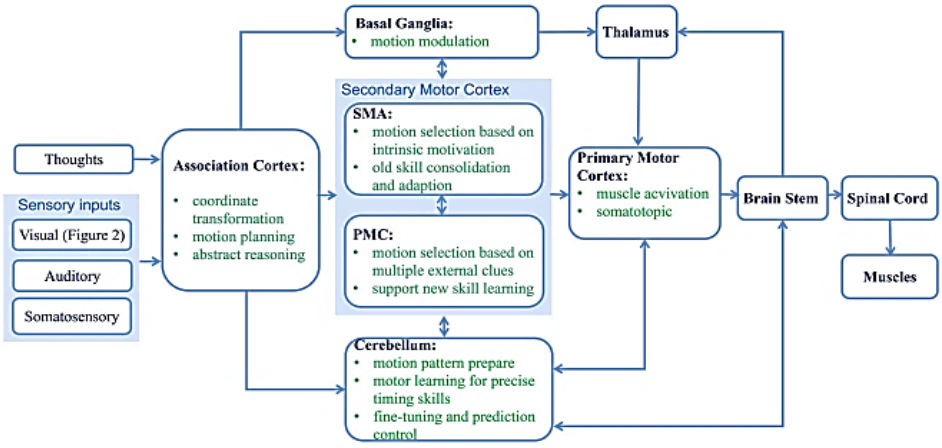
الشكل 3-3 خريطة (homunculus) (Sinha et al,2019).

3.2.2.1. المسار الحسي الحركي:

يتم التحكم في تخطيط وتنظيم الحركات اليدوية كالتخطيط المسبق واختيار الأنماط الحركية المناسبة من خلال مسارات حركية محددة في الدماغ والجهاز العصبي، وينظر إلى هذه المسارات على أنها بنية هرمية، حيث يمر التحكم الحركي من مستويات مختلفة من المستويات العليا في الدماغ (كالقشرة المخية) إلى المستويات السفلى (كالحبل الشوكي). يشمل المسار الحسي الحركي القشرة الترابطية وبشكل رئيسي كلا من القشرة الجدارية الخلفية (Posterior Parietal Cortex - PPC)، وقشرة الفص الجبهي الأمامية الظهرية الجانبية (Dorsolateral Prefrontal Cortex – DLPFC)، وتشارك هذه المناطق في تحويل الإحداثيات بين أجزاء الجسم، وتخطيط الحركة، والاستدلال المجرد، وغيرها. يوضح الشكل (3-4) رسما مبسطا للنظام الحسي الحركي، وتحديدا تقوم القشرة PPC بدمج معلومات حسية متعددة، وتعمل كواجهة حسية حركية لتخطيط الحركة والتحكم الفوري في حركات الذراع واليد الموجهة بصريا. بعد ذلك، يوجد دائرتان منفصلتان لتخطيط الحركة:

1) **الدائرة الخارجية:** تشمل القشرة PMC والمخيخ والقشرة الجدارية، وتدعم تعلم المهارات الجديدة.
2) **الدائرة الداخلية:** تشمل المنطقة الحركية التكميلية (Supplementary Motor – SMA) وقشرة الفص الجبهي والعقد القاعدية تفضل ترسيخ المهارات القديمة وتكييفها.

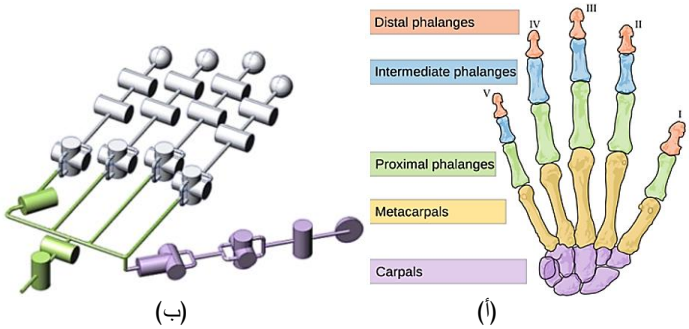
في النهاية تُرسل مخرجات هذه المناطق إلى القشرة الحركية من أجل تفعيل العضلات. إضافة لذلك، يمكن أن يوفر المخيخ تغذية راجعة تنبؤية حسية تسمح بالتحكم التنبؤي. وتقوم العقدة القاعدية بتعزيز اختيار الإجراء الأفضل وهو أمر ضروري لاكتساب التسلسلات الحركية الجديدة في مراحلها الأولى. كما تشارك مناطق مختلفة في القشرة PMC بالتحكم بإمساك الأشياء كالمنطقة (Frontal - F6) تتحكم في موعد وكيفية (نوع قبضة) الإمساك (Li et al, 2022).



الشكل 4–3 رسم تخطيطي مبسط للنظام الحسي الحركي. توضح النصوص الخضراء في كل كتلة وظائف المنطقة المقابلة (Li et al, 2022).

3.2.2.2. الهيكل العظمي والعضلات والأوتار:

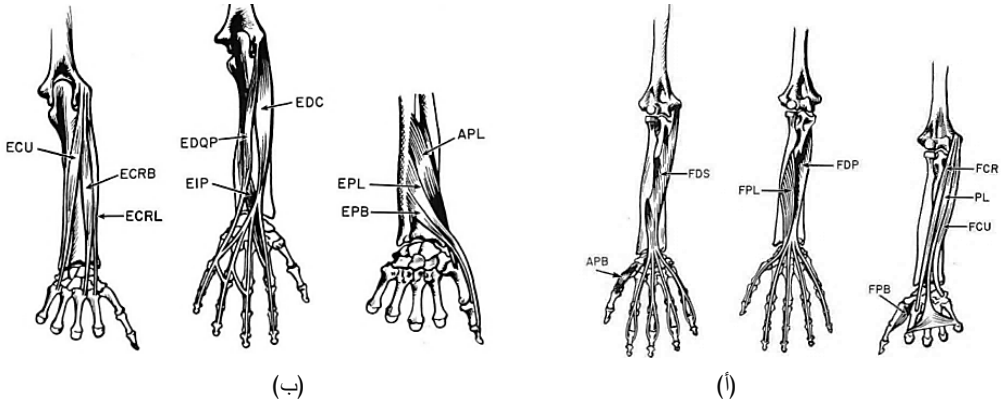
يساهم التنسيق بين الهيكل العظمي والعضلات والأوتار في اليد بإتمام الحركات الدقيقة كالإمساك والوصول للأشياء. **الهيكل العظمي:** تتكون الذراع البشرية من ثلاث عظام، عظمة الذراع العلوية (العضد) وعظمتان في الساعد (الزند، الكعبرة)، وتشكل هذه العظام 6 مفاصل تشمل المفصل (القصي الترقوي، الأخرمي الترقوي، الكتف، الكوع، الكعبري الزندي، والرسغ (المعصم)). مما يسمح بسبع درجات من الحرية في الحركة. أما اليد البشرية فلديها هيكل عظمي أكثر تعقيدا لأداء حركات دقيقة ومعقدة، حيث تتكون من 27 عظمة تشكل 15 مفصلا. وكما هو موضح في الشكل (3–5 (أ)) كل إصبع يتكون من ثلاث عظام، السلاميات البعيدة، والمتوسطة، والقريبة. تتصل السلاميات البعيدة بالقريبة عبر المفاصل بين السلاميات، وتتصل السلاميات القريبة بعظام المشط اليدوي عن طريق المفصل السلامي السنعي. وتلتقي اليد مع الرسغ بمفاصل بين الرسغ، والمفاصل شبه السنعية المنحرفة (Trapeziometacarpal – TMC) أو مفصل الرسغ السنعي (Carpometacarpal – CMC). تملك اليد 24 درجة من الحرية تشمل درجات مفصل الرسغ، يوضحها الشكل (3–5 (ب)) حيث أن كل إصبع يحتوي على أربع درجات حرية ويُنسب إلى الإبهام إما أربع أو خمس درجات حرية. يمتاز الهيكل العظمي لمفصل السلاميات المشطية (Metacarpophalangeal – MCP) بخصائص رائعة تسمح للأصابع بأداء حركات التقريب/التبعيد (adduction/abduction) والثني/البسط (flexion/extension) في آنٍ واحد، مما يزيد من براعة النظام بالكامل. وقد استلهم هذا الهيكل المعقد والتصميم الحيوي في الكثير من تصاميم الأذرع الروبوتية، والأطراف الاصطناعية (Zecca et al, 2002).



(ب) (أ)

الشكل 5–3 اليد البشرية: (أ) مخطط يوضح عظام اليد البشرية، (ب) نموذج يوضح عدد درجات الحرية لليد البشرية (Redman, 2016).

الوحدات العضلية الوترية: تشكل عضلات الطرف العلوي نظاما معقدا للغاية، ويمكن تقسيمها إلى ثلاث فئات وهي: (عضلات الكتف، عضلات العضد التي تؤثر على الساعد، عضلات الرسغ واليد)، ويتم تصنيف العديد من هذه العضلات أيضا إلى عضلات وترية خارجية مثل (المثنيات والباسطات الطويلة)، توجد في الساعد وتتحكم في الحركات العامة (الخشنة)، وعضلات داخلية تقع داخل اليد، وتكون مسؤولة عن الحركات الدقيقة، والمرونة العالية، ومهارات كالعزف والخياطة. أما في اليد يوجد 38 وحدة عضلية وترية (داخلية وخارجية) متصلة بالعظام، تتحكم في حركة اليد. أحد طرفي هذه العضلات الخارجية يكون مثبتا في عظام الذراع، وتمتد بطول الساعد حتى تصل إلى الرسغ، حيث تتحول إلى أوتار. أما العضلات الداخلية فتوجد داخل اليد نفسها وتنقسم إلى أربع مجموعات: عضلات الإبهام وعضلات الخنصر وعضلات بين العظام وعضلات دودية الشكل. يوضح الشكل (6–3) عضلات وأوتار بسط وثني المعصم والأصابع. كذلك حاولت العديد من الأنظمة الروبوتية والأطراف تقليد هذا النظام من الأوتار والعضلات في اليد البشرية، سواء من حيث التركيب أو مصدر الطاقة (Redman, 2016).



الشكل 6–3 العضلات المحركة لليد: (أ) العضلات المثنية للمعصم والأصابع، (ب) العضلات الباسطة للمعصم والأصابع (Redman, 2016).

3.2.2.3. نمط التحكم في اليد:

التحكم اليدوي الدقيق هو عملية تحكم مستمرة تتكون من خمس مراحل:

1) نقل اليد بواسطة الذراع: وفيه تتحرك الذراع بسرعة على شكل جرس (تسارع ثم تباطؤ).
2) تهيئة اليد المسبق: تتشكل اليد بوضعية تناسب شكل الجسم الهدف ويحدث آخر 30–40 % من النقل.
3) تحديد آلية الإمساك: ويعتمد عوامل عدة مثل (إغلاق الشكل أو القوة أو الوظيفية وثبات وأمان الإمساك).
4) مرحلة التلاعب: وتشمل تنسيق حركة اليد والذراع أو التبديل بين أنماط التفاعل المختلفة بين اليد والشيء.
5) التحرير والعودة: تحرير الجسم وإرجاع الذراع واليد إلى وضع الراحة.

أثناء هذه العملية تبدي الخلايا الحسية العميقة والقشرة الحركية تفضيل قوي لوضعيات اليد المتغيرة بمرور الوقت أثناء مهام الإمساك على النقيض من تفضيلها الشديد للسرعات المتغيرة أثناء مهام الوصول (Li et al, 2022).

3.2.2.4. ترميز الحركة:

يشير الترميز إلى طريقة الدماغ في تنظيم المعلومات الحركية الواردة من مجموعات من الخلايا العصبية. أشهر أنواعه يُسمى متجه التجمع العصبي (Population Vector) وهو نموذج حسابي يعتمد على فكرة أن كل خلية عصبية في القشرة الحركية تستجيب بشكل أفضل لاتجاه معين من الحركة (preferred direction) ويمكن استخدام هذه الاستجابات للتنبؤ باتجاه الحركة الفعلي ويُستخدم كثيرا في واجهات الدماغ الآلة (Brain–Machine Interface - BMI) ويُظهر اتجاه الحركة. بالنسبة لمهمة التلاعب باليد وفي مهام الإمساك التي تعمل كمولد لأنماط الحركة تُظهر القشرة الحركية الأولية أنماطا ديناميكية للتشفير على شكل متجه منخفض الأبعاد في الحركات البسيطة مثل الوصول لكنها تكون أعلى أبعادا وأكثر لا خطية أثناء الإمساك. يكون أحد الأسباب في جعل الإمساك أكثر تعقيدا هو أن الإمساك يعتمد على مدخلات حسية واردة أكثر من الحركات الأخرى. ويكون السبب الآخر هو أن الخلية العصبية الواحدة في القشرة الحركية تتحكم في أكثر من إصبع بل وأحيانا تُنشط وتُثبط العديد من العضلات معا. وبالتالي، يمكن استنتاج أن عضلات اليد يتم التحكم فيها بمستوى من التنسيق متعدد المفاصل (Li et al, 2022).

3.2.2.5. التحكم عبر تنسيق المفاصل المتعددة:

يشرح التنسيق متعدد المفاصل إلى آلية للدماغ في تنسيق حركة العديد من المفاصل في نفس الوقت لضمان حركة سلسة مثل الوصول لهدف أو الإمساك بشيء. بشكل عام لا يتم التحكم بكل عضلة في اليد بشكل مستقل بل يتم التحكم بها من خلال تنسيقات جماعية متعددة المفاصل، وبالمقارنة مع طريقة التحكم المستقلة فإن طريقة التحكم هذه تعتبر مستوى متقدم من مستوى التحكم الهرمي ويمكن أن تفيد في توليد وتعلم مهارات جديدة وتبسيط التخطيط. يتوافق هذا التنسيق مع وضع التحكم الموجود في الجهاز العصبي المركزي والذي يُطلق عليه عادةً اسم التآزر العضلي، ويُعرَّف بأنه التنشيط المتماسك لمجموعة من العضلات في المكان والزمان. بالنسبة للإمساك باليد والتلاعب بها، يتم تحليل التآزر العضلي المكاني والزماني لدى البشر من خلال تطبيق طرق التحليل، مثل تحليل المكونات الرئيسية (Principal Component Analysis – PCA)، يمكن تحقيق التحكم من خلال الفضاء الفرعي منخفض الأبعاد الناتج بحيث يمكن أن تفسر 2–4 تآزرات تباينا بنسبة 80–95% في حركات اليد (Li et al, 2022).

3.2.2.6. التحكم الهرمي والاستقلال الجزئي:

بالحديث عن البنية الهرمية للتحكم الحركي وعلى الرغم من التعقيد في التغذية الراجعة والاتصالات بين مناطق الدماغ المختلفة كما يوضح الشكل (4–3) والتي يمكن أن تكون مفيدة لتعديل الحركة الأفضل والتكيفي، إلا أن بعض المناطق – خاصة السفلية – تتمتع بنوع من الاستقلالية والقدرة على التحكم الموزّع. هذه المناطق يمكنها توليد حركات بسيطة متكررة دون الحاجة إلى تخطيط من المستوى الأعلى. وينطبق هذا بشكل خاص على وحدات التحكم والمؤثرات منخفضة المستوى والتي يمكنها توليد حركات بسيطة ومتكررة وقابلة لإعادة الاستخدام بسرعة دون

مدخلات الاستشعار والتخطيط من المستويات العالية. ويتضح ذلك من خلال التجارب التي يمكن فيها للقرود المشي والتسلق والتقاط الأشياء عندما تكون العقد الجذرية الظهرية الثنائية (Bilateral Dorsal Root Ganglia- DRGs) تالفة، وتَمَكُّن من يفتقر إلى القدرة على الحس العميق من تشكيل إيماءات مختلفة في بيئة مظلمة (Li et al, 2022).

3.3. الإمساك وتصنيفه:

3.3.1. تعريف الإمساك:

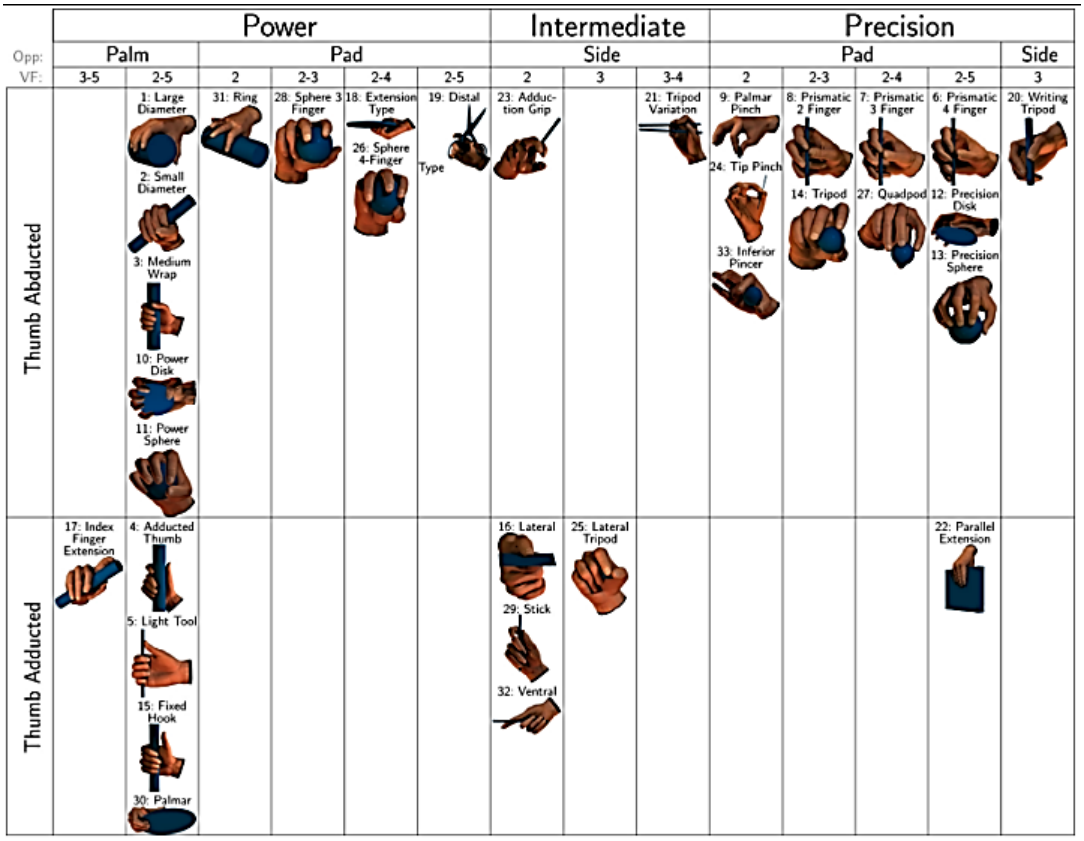
يعرف الإمساك بأنه نظام يتم بموجبه إمساك الجسم المرغوب به بيد روبوت آلي أو يد بشرية وقد يشير أحيانا إلى الإمساك بالتلاعب. وللإمساك أربع خصائص مستقلة وهي البراعة والتوازن والاستقرار والسلوك الديناميكي الذي يلزم للقيام بذلك (Kepes, 2017).

3.3.2. تصنيف الإمساك:

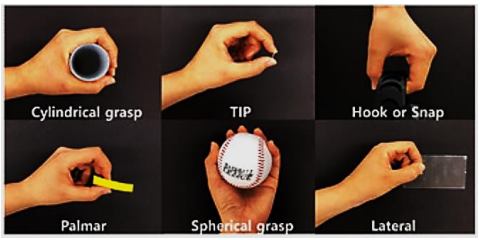
إن الدراسات السلوكية حول تصنيف الإمساك فيما يتعلق بحركة اليد وقيود كل قبضة وأنماطها تكون ذات مغزى للتعرف على نشاط اليد وإعادة التأهيل والميكانيك الحيوي. ومن أكثر التصنيفات اتساقا وكمالا التصنيف الذي نشر عام 2015 تحت عنوان "تصنيف الإمساك" المعتمد على مشروع GRASP والذي ميز 33 نوعا مختلفا من القبضات يمكن تقليصها إلى 17 فئة باستبعاد شكل / حجم الكائن وهي (Feix et al, 2016):

- فئة وفق جودة القبضة من حيث القوة والدقة والتوسيط توضح في أعمدة السطر الأول للشكل (7–3).
- فئة من النوع المتعاكس: والذي يأخذ بعين الاعتبار التقابل بين أسطح اليد موضحة في أعمدة السطر الثاني للشكل (7–3) وتصنف حسب اتجاهها إلى:
  - Pad: على طول اتجاه مواز لراحة اليد.
  - Palm: على طول اتجاه عمودي راحة اليد.
  - Side: على طول اتجاه عرضي لراحة اليد.
- فئة الإصبع الافتراضي موضح في أعمدة السطر الثالث للشكل (7–3) ويشير الأصبع الافتراضي إلى مجموعة فرعية من الأصابع عندما تعمل معا كوحدة وظيفية وتطبق قوى مماثلة لتحقيق مهمة معينة.
- فئة موضع الإبهام موضح في صفوف الشكل (7–3) تقريب أو تبعيد.

كما أن هناك تصنيف آخر قسم القبضات إلى ست فئات وهي: الأسطوانية (القوة) (cylindrical (power، والكروية spherical، والخطاف hook، والجانبية lateral، والراحية palmar، والطرفية Tip. موضحة في الشكل (8–3). ويتم تعيين القبضة الأنسب تبعا لخصائص الكائن (Li et al, 2022).



الشكل 7–3 تصنيف أنواع القبضات تبعا لم ذكره (Kepes, 2017). يوجد 17 آلية إمساك والاختلاف الرئيسي بين الإمساك داخل نفس الخلايا يرجع إلى شكل الكائن الذي يتم الإمساك به وليس أسلوب الإمساك.



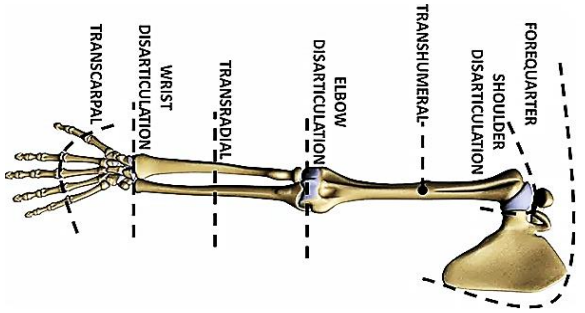
الشكل 8–3 توضيح لقبضات اليد التي وصفها Schlesinger (Song et al, 2021).

3.4. البتر والأطراف الاصطناعية العلوية:

3.4.1. البُتر والأطراف الاصطناعية:

الطرف الاصطناعي (prosthesis) هو جهاز مثبت مباشرة على جسم شخص سليم أو شخص يعاني من إعاقة حركية أو مضاعفات (مثل شخص مبتور، بحيث يحل محل جزء مفقود من الجسم والذي قد يُفقد بشكل كامل أو جزء منه نتيجة لصدمة أو مرض أو ظروف خلقية) بهدف التعويض أو زيادة قدرات الأصحاء، وتحتوي هذه الأجهزة على مشغلات وأجهزة استشعار متعددة سواء داخل جسم الإنسان أو خارجه وتستخدم هذه المعلومات الحسية الحركية لتفسير نية المستخدم وتشغيل مفاصل الطرف الآلي وفقا لذلك (Sherstan, 2015). تتم إعادة تأهيل مبتوري الأطراف الاصطناعية في المقام الأول من قبل أخصائي الأطراف الاصطناعية وفريق متعدد التخصصات من المتخصصين في الرعاية الصحية بما في ذلك الأطباء النفسيين والجراحين والمعالجين الفيزيائيين والمعالجين المهنيين ومهندسي

الطب الحيوي والبرمجيات. قد يشمل بتر الطرف السفلي إصبع القدم أو القدم أو جزء من الساق أسفل الركبة أو أعلاها أو الساق بأكملها (عند الورك) وقد يمتد البتر حتى فوق الورك. كما يشمل بتر الطرف العلوي إصبعا واحدا أو أكثر أو يدا أو جزءا من الذراع أسفل الكوع أو أعلاه أو ذراعا كاملة (عند الكتف) (Evdokia et al, 2016). يمكن تصنيف بتر الطرف العلوي إلى سبع أصناف؛ بتر اليد الجزئي وبتر الرسغ (انفصال عظمتين عند المفصل) وبتر الكوع وبتر عبر الكعبرة (بتر أسفل الكوع) وبتر عبر العضد (بتر أعلى الكوع) وبتر الكتف وبتر الربع الأمامي موضحة في الشكل (9–3). بشكل عام، كلما كان الجذمور أطول وكلما زاد عدد المفاصل التي يتم الحفاظ عليها سليمة كان من الأسهل تركيب أطراف اصطناعية لمبتور الطرف. إذا كان جزء من الجسم مفقودا غالبا ما يُنصح بطرف اصطناعي لاستبدال ذلك الجزء. كحد أدنى، يجب أن يمكّن الطرف الاصطناعي المستخدم من أداء الأنشطة اليومية (مثل المشي والأكل واللباس) بشكل مستقل ومريح بالإضافة إلى القدرة على تحمل التكاليف والتخصيص. ومع ذلك، قد تمكن الأطراف الاصطناعية المستخدم أيضا من العمل بشكل جيد أو تقريبا كما كان قبل البتر.



الشكل 9–3 مستويات بتر الأطراف العلوية (Cordella et al, 2016).

3.4.2. الأطراف الاصطناعية العلوية وأساليب التحكم والإمساك الخاصة بها:

إن منح الروبوت بشكل عام والأطراف الاصطناعية بشكل خاص قدرات شبيهة بالقدرات البشرية لأداء المهارات الحركية بطريقة سلسة وطبيعية هو أحد أهم الأهداف في تطوير الروبوتات، وإحدى الطرق الواعدة لذلك هو إنشاء روبوتات يمكنها تعلم المهارات الجديدة بنفسها على غرار البشر، ورغم ذلك فإن اكتساب المهارات الحركية الجديدة ليس بالأمر السهل وينطوي على أشكال مختلفة للتعلم، وعلى مر السنين تطورت الأساليب المتبعة لتعليم الروبوتات مهارات جديدة، يوجد الآن ثلاث أنواع رئيسية لهذا الهدف موضحة في الجدول (1–3) وهي (Kormushev et al, 2013):

1. **<u>البرمجة المباشرة:</u>** النهج الأدنى مستوى، الذي لا يزال مستخدما بنشاط في الإعدادات الصناعية، حيث تكون البيئة منظمة بشكل جيد والتحكم الكامل في الحركة الدقيقة للروبوت أمر بالغ الأهمية.
2. **<u>التعلم بالتقليد:</u>** يطلق على هذا النهج اسم التعلم بدلا من البرمجة، وذلك للتأكيد على الدور النشط الذي يلعبه الوكيل (الروبوت) في العملية. يُعرف هذا النهج أيضا باسم البرمجة بالعروض (Billard et al, 2008) أو التعلم من العروض (Argall et al, 2009). هناك ثلاث طرق رئيسية تُستخدم لإجراء العروض التوضيحية للتعلم بالتقليد:

- التدريس الحركي: فيه تتم عملية تحريك جسم الروبوت يدويا وتسجيل حركته، ينجح هذا النهج فقط مع الروبوتات الصغيرة وليس عمليا بالنسبة للروبوتات الأكبر حجما.
- التشغيل عن بعد: يشترك هذا النهج في العديد من أوجه التشابه مع التدريس الحركي، وأحد الفروق الكبيرة هو أنه مع التشغيل عن بعد يمكن أن يكون المعلم موجودا في مكان جغرافي بعيد. ومع ذلك، تصبح تأخيرات الوقت مشكلة مع ازدياد المسافة، ويكون من الصعب الشعور بالقيود وقدرات الروبوت مقارنة باستخدام التدريس الحركي. وعلى غرار التدريس الحركي، لا يمكن التحكم إلا في عدد محدود من درجات الحرية في نفس الوقت. كميزة لهذا النهج، يمكن أن يكون الإعداد أبسط في بعض الأحيان.
- التعلم بالملاحظة: فيه توضح الحركة باستخدام جسم المعلم نفسه ويتم إدراكها باستخدام أنظمة التقاط الحركة أو كاميرات الفيديو أو أجهزة الاستشعار الأخرى. عادة ما يكون ذلك ضروريا لحل مشكلة المراسلات، أي لربط حركيات الروبوت بحركيات المعلم. إن أبسط طريقة لاستخدام العروض التوضيحية هي القيام بتسجيل وإعادة تشغيل بسيطين والذي لا يمكنه تنفيذ سوى مثيل معين من المهمة. ومع ذلك، باستخدام التمثيل الذكي للحركة المسجلة في إطار مرجعي مناسب، فمن الممكن أن يكون لدى الروبوت مهارة قابلة للتكيف إلى حد ما مع التكوينات الأولية المختلفة، للمهام البسيطة. بناء على عروض توضيحية متعددة تتضمن اختلافات في المهمة، يمكن للروبوت أن يحسب الارتباطات والتباين ويحدد أي جزء من المهمة من المهم تكراره حرفيا وأي جزء مقبول للتغيير وإلى أي مدى.

3. **<u>التعلم الذكي (التعلم المعزز):</u>** أنشأ هذا التعلم مكانة مهمة جدا لتطبيقه في مجال الروبوتات وهو عملية التعلم من التجربة والخطأ أثناء استكشاف البيئة وجسم الروبوت نفسه. يتم تحديد الهدف فيه من خلال دالة مكافأة تعمل كمعزز إيجابي أو سلبي اعتمادا على أداء الروبوت فيما يتعلق بالهدف المطلوب. كما يوفر بعض المزايا الإضافية مثالا يمكن البدء بالتعليم من حالة جيدة كفاية وتحسينها تدريجيا. كذلك له القدرة على التكيف ديناميكيا مع التغييرات في الوكيل نفسه كالروبوت الذي يتكيف مع تغييرات الأجهزة من تسخين وتآكل ميكانيكي ونمو أجزاء الجسم. والدافع الرئيسي لاستخدامه في تعليم مهارات جديدة هو أنه يوفر ثلاث قدرات مفقودة سابقا وهي:
   - تعلم مهام جديدة، والتي لا يستطيع حتى المعلم البشري إظهارها جسديا أو لا يستطيع برمجتها بشكل مباشر (على سبيل المثال، القفز لارتفاع ثلاثة أمتار، ورفع أوزان ثقيلة، والتحرك بسرعة كبيرة)؛
   - إيجاد الحلول المثلى للمشاكل الصعبة التي ليس لها صياغة تحليلية أو حل معروف، باستخدام دالة تكلفة معروفة فقط (على سبيل المثال، تقليل الطاقة المستخدمة لأداء مهمة أو إيجاد أسرع طريقة مشي)؛
   - تعلم كيفية تكييف مهارة قديمة مع مهمة جديدة غير مرئية سابقا (مثل، تعلم المشي من أرض مستوية إلى منحدر، وتعميم مهارة المشي المستوي على مهمة جديدة). يمكن لبعض أساليب التعلم بالتقليد القيام بذلك أيضا ولكن بطريقة أكثر تقييدا (كضبط محددات النموذج المكتسب دون القدرة على تغييره).

الجدول 1-3 مقارنة بين مناهج التعليم الرئيسية للروبوتات (Kormushev et al, 2013).

أسلوب تعلم الروبوت		التعقيد الحسابي / صعوبة التعلم	المزايا	العيوب
البرمجة الكلاسيكية (برمجة مفاصل الروبوت يدويا لكل حركة)		منخفض / عالي	التحكم الكامل في حركة الروبوت حتى أدنى مستوى	يستغرق وقتا طويلا، عرضة للخطأ، غير قابل للتطوير، قابل لإعادة الاستخدام
التعلم بالتقليد	التدريس الحركي والتشغيل عن بعد (تحريك أطراف الروبوت بشكل مباشر للحركة الهدف أو التحكم عن بعد لتسجيل الحركة)	متوسط / منخفض	لا حاجة للبرمجة اليدوية، يمكن فقط تسجيل الحركات وإعادة تقليدها.	لا يمكنه القيام بحركات سريعة، يمكن تحريك طرف واحد فقط في كل مرة، يجب أن يكون الروبوت خفيف الوزن
	التعلم بالملاحظة (إظهار الحركات باستخدام جسم المعلم نفسه وإدراك الحركة باستخدام أنظمة التقاط)	متوسط / قليل	يكون من السهل تقليد الحركات كما أنه يعمل مع مهام ثنائية اليدين وحتى لحركة كامل الجسم	مشكلة المراسلات الناجمة عن التجسيد المختلف، يجب أن يكون المعلم قادر على أداء المهمة، غالبا ما يتطلب الأمر عروض توضيحية متعددة تحتاج إلى تقسيمها وتنسيقها زمنيا
التعلم المعزز (فيه يوجه سلوك الوكيل من خلال دالة مكافأة، ولا يحتاج إلى توضيح كيفية أداء المهمة، إذ يكتشف الروبوت ذلك بنفسه)		مرتفعة / منخفضة	يمكن للروبوت تعلم مهارات لا يستطيع حتى الإنسان القيام بها واكتشاف أساليب جديدة لتحقيق الهدف قد تكون أكثر مثالية	لا يوجد سيطرة على التصرفات، الروبوت لديه معلومات غير مباشرة حول الهدف، الحاجة لتحديد تابع المكافأة، وضبط محددات السياسة الأولية كحجم الاستكشاف / والاستراتيجية

3.4.3. أساليب نقل خصائص الإمساك البشري إلى الأيدي الروبوتية والاصطناعية:

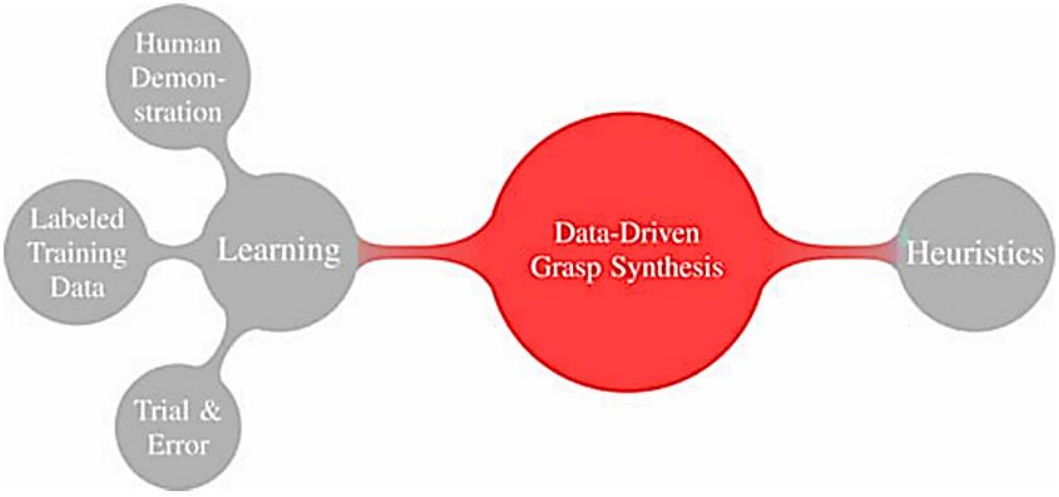
هناك أساليب متنوعة وعديدة لنقل ومحاكاة خصائص الإمساك الموجودة في الأيدي البشرية (الموضحة في الفقرة 3.1) إلى الأيدي الروبوتية، وبشكل عام يمكن تقسيم هذه الأساليب إلى نهجين هما: نهج الإمساك التحليلي، ونهج الإمساك المبني على البيانات، يضاف إليهم النهج الهجين الذي يجمع الأسلوبين معا.

3.4.3.1. النهج التحليلي والمعروف أيضا باسم النهج الهندسي:

تستند هذه الأساليب إلى الإنشاءات التحليلية لتحليل شكل الكائن وفهم كيفية تحديد وضع الإمساك المناسب. فيها يُحَقَّق الإمساك من خلال مشكلة أمثلة مقيدة تتطلب تركيبات هندسية و/أو حركية و/أو ديناميكية. ولتحقيق إمساك ناجح يجب استيفاء عدد كبير من القيود وبالتالي هي أساليب معقدة. وتركز معظم أساليب النهج التحليلي إما على نموذج اليد أو قيود المهمة والتي يشار إليها عادة بـ القوة المغلقة، أو توجيه المهام. يمكن أن تؤدي الأخطاء في النماذج الحركية والديناميكيات للروبوت والضوضاء في قيم المستشعرات إلى جعل النماذج التحليلية غير دقيقة مما يؤدي إلى زيادة صعوبة تحقيقها للإمساك. غالبا ما تستهلك الأساليب التحليلية للإمساك الآلي وقت كبير وتعاني من نقص قدرات التعميم لأشياء جديدة. سادت الأساليب التحليلية حتى عام 2000، ومع دخول AI أصبحت مناهج البيانات مطبقة على نطاق أوسع (Kepes, 2017) (Danielsen et al, 2021).

3.4.3.2. المناهج المبنية على البيانات والمعروفة باسم المناهج التجريبية:

تعالج هذه الأساليب صعوبات النهج التحليلي. وتتجنب الصياغات الرياضية والفيزيائية المعقدة وتعتمد على تعلم خبرات إمساك العينات من قاعدة بيانات تم إنشاؤها بواسطة عروض بشرية أو طرق أخرى وترتيبها وفقا لبعض المقاييس التجريبية. يستخدم فيها المعلومات الإدراكية كاستخراج السمات والدلالات والتصنيف وغيرها لحل مهمة إمساك معينة. يمكن تطوير تقنية إمساك من خلال التعلم والتعميم من البيانات أو باستخدام بعض التجارب والتعلم منها كما يوضح الشكل (10–3). عادةً يتم إنشاء قاعدة البيانات المستخدمة للتعلم من خلال العرض التوضيحي البشري والتجربة والخطأ أو يُفترض أنها مجموعة بيانات مصنفة. تتمثل المزايا الرئيسية لهذه الحلول في تعقيد حسابي أقل، وافتراضات أقل، وهشاشة أقل، بينما تتطلب عادة كمية كبيرة من البيانات. اكتسبت الأساليب القائمة على البيانات شعبية في السنوات الأخيرة بسبب زيادة توافر مجموعات البيانات والموارد الحسابية والخوارزميات الأقوى وتستخدم تعلم الآلة والتعلم العميق (Danielsen et al, 2021) (Kepes, 2017).



الشكل 10–3 نهج الإمساك المقاد بالبيانات (Bohg et al, 2013).

3.5. الخاتمة:

لقد تم البحث في التحكم في الأذرع الاصطناعية التي تعمل بالطاقة لأكثر من 50 عاما، ومع ذلك، فإن التحكم في الأطراف الاصطناعية لا يزال مشكلة مفتوحة، ليس فقط من منظور بحثي، ولكن أيضا من منظور سريري. لقد تم تحقيق تقدم كبير في تصنيع الأطراف الاصطناعية عالية الأداء، ومع ذلك، فإن التحكم في هذه الأطراف لا يزال غير عملي إلى حد كبير. تكمن القضية الأساسية في غياب التطابق الكبير بين عدد الوظائف المتاحة في الأذرع الاصطناعية الحديثة التي تعمل بالطاقة وعدد الوظائف التي يمكن لمبتور أن يقوم بها بنشاط في أي لحظة. أحد الأساليب لمعالجة هذا التباين هو فكرة التعامل مع الذراع كوكيل ذكي يسعى إلى تحقيق هدف معين يمكن لمثل هذا الوكيل أن يتعلم من تجربته ويكيف أفعاله لتحسين قدرته على تحقيق الهدف. يُفترض أن مثل هذه العوامل الذكية ستكون قادرة على التعويض عن القيود الحالية بين الذراع الاصطناعية المزودة بالطاقة والشخص المبتور.

4. الفصل الرابع: الأسس النظرية للبحث.

4.1. تمهيد:

في هذا الفصل تُقدم مراجعة عن التعلم المعزز والتعلم المعزز العميق وآلية تطبيقهم في حل مسائل الإمساك بواسطة يد آلية متعددة الأصابع، بالإضافة لأهم الأسس النظرية والتي يعتمد عليها هذا البحث.

4.2. مفاهيم مهمة:

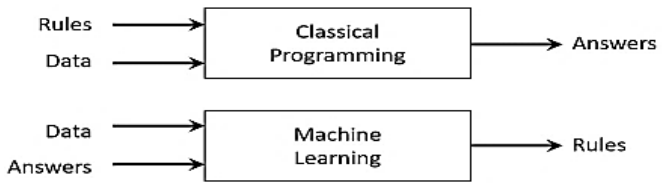
مؤخرا، أصبح التكامل التكنولوجي للذكاء الاصطناعي وتعلم الآلة في صناعة الأطراف الاصطناعية وتقويم العظام وفي مجال التكنولوجيا المساعدة نعمة لذوي الاحتياجات الخاصة، حيث تم استخدام مفهوم الشبكة العصبية من قبل الشركات المصنعة الرائدة لوسائل إعادة التأهيل لمحاكاة وظائف تشريحية وميكانيكية حيوية مختلفة للأجزاء المفقودة من جسم الإنسان. فتطبيق الذكاء الاصطناعي وتكنولوجيا الروبوتات له تأثير كبير في تحقيق الاستقلال وتعزيز نوعية الحياة. وقبل البدء بالحديث عن التعلم المعزز العميق تُعرض مجموعة من المفاهيم الأساسية.

4.2.1. الذكاء الاصطناعي (Artificial Inelegant - AI):

هو جعل الآلة تفكر وتتصرف كإنسان وتحاكي الذكاء البشري (Mahmud et al, 2018)، ويعتبر العالم الأميركي John McCarthy أول من صك مصطلح الذكاء الاصطناعي عام 1956 ميلادي، وعرّفه بأنه علم وهندسة صناعة الآلات وبرامج الحاسوب الذكية، أو هو فرع علوم الحاسوب الذي يهدف إلى إنشاء الآلات الذكية. ولاحقا لذلك عرّف بعدة تعريفات تقود جميعها لفكرة أن AI بشكل عام، هو مفهوم شامل يشير إلى أنظمة كمبيوتر قادرة على إدراك بيئتها، والمنطق، والتعلم، وإدارة البيانات حتى تتمكن من التصرف وفقا لأهدافها وما تدركه. والموضوع الأكثر بروزا فيه هو تعلم الآلة (Sekkat et al, 2023).

4.2.2. تعلم الآلة (Machine Learning - ML):

هو مجموعة فرعية من الذكاء الاصطناعي حيث تتعلم الآلة أداء المهام الغير مخطط لها بشكل صريح، يحوي أدوات رياضية وإحصائية وعلوم الحاسوب، ما يدفع عجلة التقدم في تطوير الذكاء الاصطناعي (Mahmud et al, 2018). يختلف تعلم الآلة عن البرمجة الكلاسيكية في كيفية الوصول لقواعد البرنامج أو منطقه، ففي البرمجة الكلاسيكية يحدد المبرمج مجموعة من القواعد ويكتب برنامج يأخذ بيانات الإدخال ويعالجها وفقا لهذه القواعد ثم يعطي الإجابات. في حين يعمل تعلم الآلة على أتمتة عملية الوصول للقواعد حيث يأخذ بيانات الإدخال والإجابات المطلوبة ويخرج القواعد، ويوضح الشكل (1-4) هذا الفرق (Danielsen et al, 2021).



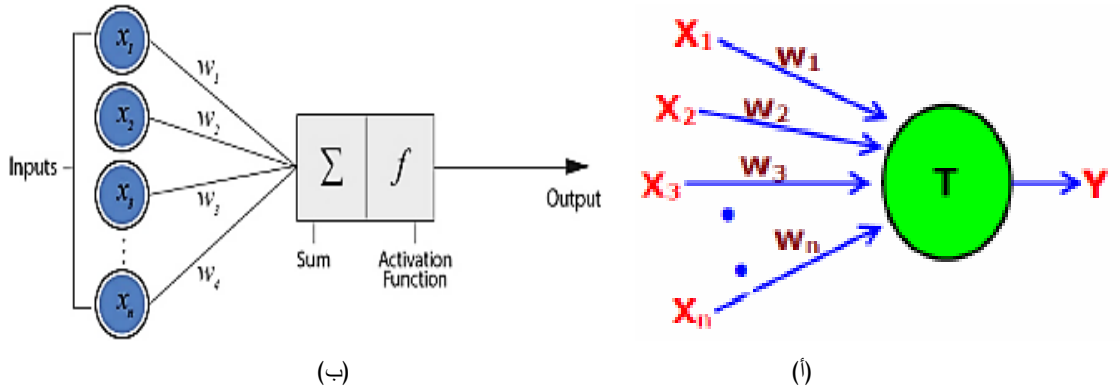
الشكل 1-4 تعلم الآلة مقابل البرمجة الكلاسيكية (Danielsen et al, 2021).

وبشكل عام يكون التصنيف الأكثر شيوعا لخوارزميات ML هو تصنيفها إلى ثلاثة أنواع للتعلم وهي: التعلم الموجه، والتعلم غير الموجه، والتعلم المعزز. ويتم تمييز نماذج التعلم هذه فيما بينها من خلال نوع البيانات التي يتعلم النظام منها. ففي التعلم الموجه يتم توفير مجموعة تدريب مصنفة من أزواج (مدخلات، مخرجات). تتكون المدخلات في عينة مجموعة التدريب من بعض الميزات والخصائص للتعلم منها، والمخرجات المعروفة أيضا باسم (label) هي الاستجابة الصحيحة للإدخال المحدد. تتدرب خوارزمية التعلم الموجهة على العينات في مجموعة التدريب وتحاول تعلم العلاقة بين المدخلات والمخرجات عن طريق نموذج محدد يُخطط من المدخلات إلى المخرجات. الهدف من التعلم الموجه هو التعميم على عينات جديدة لم يصادفها في مجموعة التدريب. يمكن تجميع خوارزميات التعلم الموجه في خوارزميات التصنيف والانحدار. يختلف التعلم غير الموجه عن التعلم الموجه، حيث أنه يكون لديه بيانات لكنها غير مصنفة، بمعنى لا يوجد label موفرة لخوارزمية التعلم. تحاول الخوارزمية اكتشاف أوجه التشابه والأنماط في بيانات الإدخال. المهام الأكثر شيوعا للتعلم غير الموجه هي: التجميع والحد من الأبعاد. يكمن التعلم المعزز في مكان ما بين أساليب التعلم الموجه وغير الموجه. في التعلم المعزز، تتلقى خوارزمية التعلم تعليقات ما إذا كانت الإجابة صحيحة أو خاطئة، لكنها لا تحصل على أي تعليمات حول كيفية تصحيح الإجابة، فتستكشف الخوارزمية إمكانيات مختلفة من خلال التجربة والخطأ حتى تكتشف كيفية تحقيق الهدف (Danielsen et al, 2021).

4.2.3. الشبكات العصبية الاصطناعية والخلايا العصبية الاصطناعية:

الشبكات الاصطناعية هي تقنية من تقنيات تعلم الآلة مستوحاة من الشبكات العصبية البيولوجية. وتتكون من عدد كبير من الخلايا العصبية والوصلات بينها.

الخلايا العصبية الاصطناعية: هي اللبنة الأساسية للشبكات العصبية. وهي وحدة رياضية تأخذ المجموع المرجح لمدخلاتها، وتطبق دالة التنشيط عليها و"تنشط" الخلايا العصبية إذا كان المجموع المرجح أكبر من عتبة معينة. أبسط أمثلتها البيرسبيترون وهو نموذج من الشبكات ذات العصبون الوحيد قابلة للتعلم موضحة في الشكل (2-4(أ)) وبشكل أكثر عمومية يكون للعصبون الاصطناعي الهيكل الموضح في الشكل (2-4(ب)) ضمن الشبكة (Kepes, 2017) (www.myreaders.info).



(ب) (أ)

الشكل 2-4 العصبون الاصطناعي

(أ) نموذج بيرسبيترون (www.myreaders.info). (ب) الشكل العام للعصبون الاصطناعي ضمن الشبكة الاصطناعية (Kepes, 2017).

- $X_1, X_2, \ldots, X_n$: هي المدخلات كأعداد حقيقية أو منطقية اعتمادا على نوع المسألة.
- $W_1, W_2, \ldots, W_n$: هي أوزان ذات قيم حقيقية تعطى بحسب أهمية المدخلات.
- T: هي العتبة وتكون قيم حقيقية.
- Y: هي مخرجات البيرسبيترون وتكون ذات قيم منطقية.
- إذا كانت نتيجة $X_1W_1 + X_2W_2 + \ldots + X_nW_n$ أكبر من T ستكون $Y = 1$، وإلا فستساوي0. توضح المعادلة (4-1) الصيغة الرياضية لعمل الخلية العصبية كما يلي:

$$Output = Y = \begin{cases} 1 & if\ b + w^t x > 0 \\ 0 & otherwise \end{cases} \quad (4-1)$$

أو عند تمثل العتبة بانحياز سلبي $t = -b$ كما في المعادلة (4-2):

$$Output = Y = \begin{cases} 1 & if\ w^t x > t \\ 0 & otherwise \end{cases} \quad (4-2)$$

دالة التنشيط أو التفعيل: تستخدم ضمن الشبكات العصبونية لإجراء عمليات حسابية معقدة في الطبقات الخفية. ثم تنقل النتيجة إلى طبقة الخرج. يوجد الكثير منها إلا أن عدد قليل شائع الانتشار مثل:

- الخطية Linear: يأخذ العصبون فيها مجموعة مدخلات ويضيف قيمة الانحياز b ثم يخرج الرقم الناتج وتعطى بالعلاقة: $a = g(a)$
- السينية Sigmoid: أكثر التوابع شيوعا، تقوم بتقييس المدخلات ضمن [0,1]، وتعطى بالعلاقة: $g(a) = \frac{1}{1+e^{-a}}$
- الظل الزائدي Tanh: لها أيضا نطاق محدد ضمن [1,−1]، وتعطى بالعلاقة: $g(a) = \frac{e^{2a}-1}{e^{2a}+1}$
- الخطي المصحح ReLU: تابع خطي يأخذ القيم الموجبة ويصفر السالبة: $g(a) = \max(0, a)$

إن لاختيارها أثر كبير على النتائج، وعادة تستخدم طبقة الخرج دالة تنشيط مختلفة عن دالة تنشيط الطبقات الخفية.

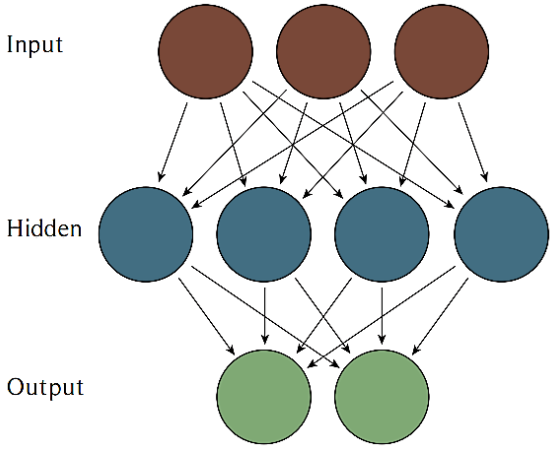
ويصبح الآن تمثيل الخلية العصبية رياضيا كدالة كالتالي:

$$h(x) = g\big(a(x)\big) = g(b + w^T x) \quad (4-3)$$

حيث: $h(x)$ هي الخلية العصبية كدالة، وg تابع التنشيط، وa المجموع المرجح وانحياز الدخل.

<u>الشبكة العصبية الاصطناعية:</u> هي نماذج تتدرب من أجل التعرف على الأنماط الأنسب للبيانات، بحيث تأخذ شعاع أو متجه من القيم كمدخلات وتعطي متجه من القيم كمخرجات، أحد الهياكل الناجحة لهندسة الشبكة العصبية هو الهيكل الذي يتم من خلاله تقسيم الخلايا العصبية إلى طبقات تكون بينها وصلات كما يوضح الشكل (3−4). عادة ما تكون الطبقات المتجاورة متصلة تماما، وهناك اتصال بين كل خليتين عصبيتين من طبقتين متجاورتين.

تسمى هذه الأنواع من الشبكات العصبية شبكات التغذية الأمامية. وتسمى الطبقات الوسيطة الطبقات المخفية لأننا لا "نرى" مدخلاتها ومخرجاتها. في الشبكات متعددة الطبقات تكون مخارج الطبقات السابقة هي مدخلات للطبقات المخفية اللاحقة ولطبقة الخرج. إن الطبقات يمكن أن يكون لها دوال تنشيط مختلفة داخل شبكة واحدة. بالنسبة لمهام التصنيف، عادة ما يكون عدد الخلايا العصبية في طبقة الخرج هو عدد الفئات التي من المفترض أن يتم تصنيف المدخلات إليها. يمكن تقييس خرج التنشيط عبر جميع الخلايا العصبية الناتجة، وبالتالي يمكن تفسير الناتج على أنه توزيع احتمالي. قدرات الشبكة العصبية تزداد مع زيادة عدد الطبقات المخفية فمن المعروف أن خلية عصبية واحدة يمكنها تصنيف المشكلات القابلة للفصل خطيا والشبكات العصبية بطبقة واحدة مخفية قادرة على نمذجة أي وظيفة مستمرة في حين أن الشبكات متعددة الطبقات قوية بما يكفي لتمثيل أي وظيفة عن طريق تعديل فضاء الدخل بحيث يصبح قابلة للفصل خطيا (Kepes, 2017).



الشكل 3-4 شبكة تغذية أمامية مع طبقة مخفية واحدة (Kepes, 2017).

4.2.4. خوارزمية تدريب الشبكة العصبية الاصطناعية:

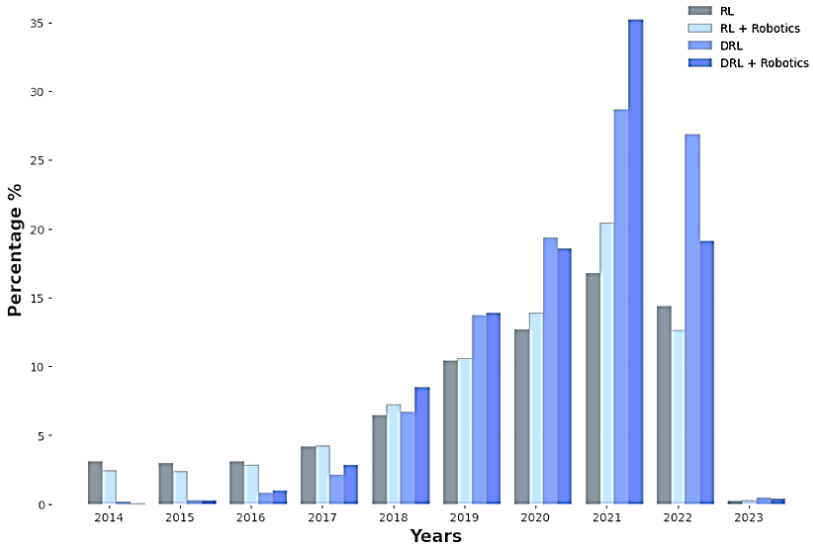
إن التحدي الرئيسي في برمجة الشبكة العصبية هو إيجاد الأوزان المناسبة للشبكة. الخوارزميات المستخدمة للقيام بذلك تسمى خوارزميات التدريب. عند تدريب شبكة عصبية بطريقة موجهة، يتم تزويدها بمجموعة كبيرة من أمثلة التدريب التي تتكون من أزواج المدخلات والمخرجات. أثناء التدريب، يقاس الفرق بين المخرجات الفعلية للشبكة والمخرجات المتوقعة من إشارة الإشراف. يشار إلى هذا القياس باسم تابع الخسارة أو الخطأ أو التكلفة (من أمثلته مجموع مربعات الخطأ، ودالة الخطأ اللوغارتمي) والهدف هو تقليله. كلما كان الخطأ أصغر، كلما تعلمت الشبكة بشكل أفضل سمات مجموعة التدريب. وطالما أن دالة الخسارة ليست صغيرة بما يكفي، فيجب الاستمرار في ضبط الأوزان. بالنسبة للشبكات ذات الطبقات المخفية، يتم استخدام خوارزمية الانتشار العكسي Backpropagation من طبقة الخرج إلى طبقة الدخل لهذه المهمة حيث تقوم الشبكة بالاعتماد على الخطأ أثناء الانتشار الخلفي لتعديل الأوزان وحساب الوزن الجديد، وتعتمد سرعة التدريب ومدى تغيراته على معدل التعلم α الذي يحدد إما يكون التدريب سريع ويحدث تغيرات كبيرة عندما يكون له قيمة عالية، أو العكس.

4.3. التعلم المعزز (Reinforcement Learning – RL):

من المعروف بالنسبة لعلم النفس والأعصاب أن الدماغ يملك أنظمة متميزة قادرة على اتخاذ القرار، وعلى الرغم من أننا بعيدون جدا عن فهم هذه الأنظمة بشكل تام، فلقد بينت التجارب أننا قادرون على نمذجة بعض الوظائف الدماغية المرتبطة بالتعلم باستخدام مفهوم التعلم المعزز (Sutton and Barto, 2018).

4.3.1. نبذة تاريخية عن RL:

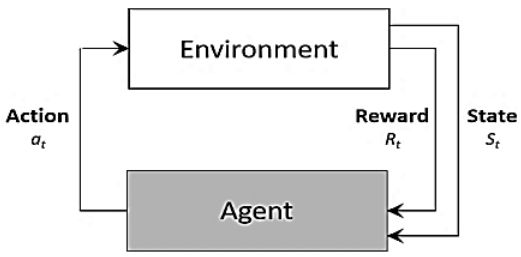
أظهر التعلم المعزز إمكانات كبيرة في مجال الروبوتات وخاصة في الإمساك بالأشياء، ويُعتبر النهج الأفضل لبناء روبوتات ذكية. يعتمد تاريخيا على منطقتين من البحث تمت متابعة كل منهما على حدا قبل إنتاج مفهوم التعلم المعزز الحديث، وهما: علم نفس الحيوان، والتحكم الأمثل. دفع علم نفس الحيوان لتطوير مفهوم التعلم بالتجربة والخطأ. حيث قُدمت نظرية التعلم بالتجربة والخطأ لأول مرة من قبل عالم النفس الشهير إدوارد إل ثورندايك ثم نفذ هذا المفهوم في بعض أعمال الذكاء الاصطناعي المبكرة وأخيرا أدى لنهضة التعلم المعزز في أوائل الثمانينات (Sun et al, 2021) (Sekkat et al, 2023). أما مشكلة التحكم الأمثل تم ابتكارها لتصميم وحدة تحكم تقلل دالة الخسارة للنظم الميكانيكية مع الوقت. في منتصف الخمسينيات من القرن الماضي، وبشكل أكثر دقة في عام 1957، قدم ريتشارد بيلمان منظورا مبتكرا لنظرية هاملتون–جاكوبي، الذي ابتكر نهجا لمعالجة مشكلة التحكم الأمثل المعروفة باسم البرمجة الديناميكية. ثم ظهرت مفاهيم أخرى تم دمجها مع المجالين السابقين في أواخر الثمانينيات. هذه المفاهيم هي مناهج الفرق الزمني ومونت كارلو وهذا الاتحاد بين المجالات الثلاثة أدى إلى ظهور مجال التعلم المعزز الحديث. وله تطبيقات واسعة في حل مشاكل التحكم والأمثلة التي تستلزم اتخاذ قرارات متسلسلة. وبالتالي، فهو موضوع ذو أهمية. يوضح الشكل (4–4) الاهتمام بهذا المجال كنسبة مئوية للأوراق المنشورة حول هذا المجال من عام 2014 حتى 2022 (Sun et al, 2021) (Sekkat et al, 2023).



الشكل 4–4 النسب المئوية للاهتمام بالتعلم المعزز وخاصة في مجال الروبوتات آخر 10 سنوات بالاعتماد على قواعد WEB OF SIENCE WOS الانخفاض الظاهر في 2023 بسبب إيقاف الدراسة التي قدمت المخطط (Sekkat et al, 2023).

4.3.2.تعريف RL:

هو نموذج لتعلم الآلة يجمع بين الشبكات العصبية التلافيفية وتخطيط بيلمان الديناميكي ويُستَلهم من نظرية التعلم السلوكي وكيف تتعلم الكائنات الحية آلية الوصول إلى هدفها وتتكيف من خلال التفاعلات مع البيئة بناء على التجربة والخطأ وقانون التأثير، وهي المفاهيم الأساسية التي يعتمد عليها RL، يهدف مفهوم قانون التأثير إلى تمكين الوكيل من اختبار الإجراءات وتلقي الملاحظات (التعزيز) ويتضمن تعديل سلوك الوكيل لتعظيم المكافأة التراكمية التي يتلقاها حيث يقترح أن الكائنات تتعلم الإجراءات أو السلوك الذي يجب اتباعه في البيئة من خلال تجربة إجراءات مختلفة مرارا وتكرارا، ويستند السلوك المستفاد إلى نوع الاستجابة التي توفرها البيئة، مما يعني أن الاستجابات التي يعتبرها الكائن مرضية تحدث في المستقبل. خوارزميات RL ليست موجهة أو غير موجهة مثل معظم خوارزميات تعلم الآلة، لأنها غير مدربة على البيانات المصنفة كما هو الحال في التعلم الموجه، يجب أن تجد أفضل الحلول من خلال الاكتشاف من خلال التجربة والخطأ، ولكن على عكس التعلم غير الموجه، لا يتعين عليها العثور على الميزات الموجودة في بيانات التدريب دون أي دليل بطريقة غير موجهة، وبدلا من ذلك يتم توجيهها بواسطة إشارة مكافأة. إذا افترض بيئة غير معروفة في البداية، فعلى الوكيل استكشافها من خلال اتخاذ قرار متسلسل يتم التحكم فيه من خلال مكافأة السلوك الصحيح ومعاقبة السلوك غير المرغوب فيه ثم اكتساب الخبرة من أجل تحسين المعرفة بهذا العالم والتفاعل معه وبالتالي تحسين وزيادة فرص البقاء. خلال هذا الاستكشاف تحصل تجارب جيدة وسيئة (أو أي شيء بينهما) وبطبيعة الحال يجب أن تعاش التجارب الجيدة مرة أخرى ويُتَعلم كيفية تحقيقها، بينما يراد تجنب التجارب السيئة. عندما تصبح الأمور أكثر تعقيدا، قد نجد أنه لا يمكن الحصول على بعض التجارب الجيدة إلا من خلال تجربة بعض التجارب السيئة أولاً. في مثل هذه الحالات يجب أن يقرر الوكيل ما إذا كانت تلك التجربة الجيدة تستحق خوض تجارب السيئة أم لا (Kepes, 2017) (Sekkat et al, (Danielsen et al, 2021) (2023. ووفقا لـ Sutton and Barto فإن RL هو تعلم كيفية الانتقال من حالة إلى أخرى لتحقيق هدف معين من خلال اختيار الإجراءات التي تعظم المكافأة. وهو مستوحى من نظريات التعلم السلوكي لدى الكائنات الحية، ويعتمد على التفاعل مع البيئة عبر التجربة والخطأ. عادة ما تصاغ هذه المشكلات كعمليات قرار ماركوف كما في الشكل (5-4)، حيث يتلقى الوكيل ملاحظات للحالة ويتخذ إجراء ويحدث تغيير في البيئة ثم يحصل على مكافأة. يهدف الوكيل لتعظيمها ولاكتشاف السياسة المثلى التي تربط كل حالة بالإجراء الأمثل (Kepes, 2017) (Gomes et al, 2012) (Sutton and Barto, 2018).



الشكل 5-4 تفاعل الوكيل مع البيئة (Danielsen et al, 2021).

يوضح الشكل (4-5) كيف يشاهد الوكيل البيئة من خلال تلقي الحالة s، والمكافأة العددية r في كل خطوة زمنية t، $0 \leq t \leq T$. ثم يقرر الإجراء التالي $a \in A$. يمكن أن يغير الإجراء البيئة ويغير حالتها التي يلاحظها الوكيل مرة أخرى. S وA هما كل الحالات وجميع الإجراءات الممكنة على التوالي وT هي طول التسلسل الزمني المحدد. لاختيار الإجراءات يستخدم الوكيل سياسة اتخاذ القرار π، والتي تحدد الحالات للإجراءات، أو توزيعات احتمالية الإجراء، وتحدد سلوك الوكيل. وكل تفاعل بين الوكيل والبيئة في لحظة زمنية t من المجال $0 \leq t \leq T$ يعرف بخطوة step. تسمى سلسلة واحدة من التفاعلات بين الوكيل والبيئة خلال زمن بالحلقة epoch أو episode.

المكافآت: يُكسب التفاعل مع البيئة الوكيل خبرة، يمكن جمع المكافأة في كل لحظة زمنية r_t للحصول على المكافأة الإجمالية لحلقة ما، وعمليا تحدد المكافأة التراكمية على أنها إجمالي المكافأة المخصومة في الخطوة الزمنية t:

$$R_t = r_{t+1} + \gamma r_{t+2} + \gamma r_{t+3}^2 + \cdots = \sum_{k=0}^{\infty} \gamma r_{t+k+1}^k \quad (4-4)$$

حيث أن $\gamma \ni [0،1]$ هي عامل الخصم، وهي معلمة فائقة تضمن أن المكافآت المستقبلية تساوي أقل من المكافآت الفورية مما يمنع الحلقات اللانهائية ويشجع الحلول القصيرة والفعالة وهو ما يشبه سلوك الإنسان والذي يفضل المكافآت الفورية. تساعد المكافآت الوكيل على تكييف سياسته التي تقترح الإجراءات التي من شأنها تعظيم المكافأة على المدى الطويل (Kepes, 2017). يوجد نوعين من توابع المكافأة، الأشهر هو تصميم وظائف مكافأة كثيفة (مُشكَّلة) Shaped (Dense) Reward، والنوع الآخر هو المكافآت المتناثرة (المتفرقة) Sparse Reward، أي مكافأة العميل فقط عند إكمال المهمة بنجاح يمكن أن يؤدي إلى سياسات أفضل ومع ذلك، فإن استكشاف فضاء الحالة والإجراء أكثر صعوبة في هذه الحالة. يعد استخدام المكافآت المتفرقة أكثر استحسانا في الممارسة العملية حيث يتم تعميمها على العديد من المهام دون الحاجة إلى الهندسة اليدوية. من ناحية أخرى، يعد التعلم بالمكافآت المتفرقة فقط أكثر تحدي نظرا لأن ربط تسلسلات الإجراءات بالمكافآت غير الصفرية التي يتم تلقيها فقط عند إكمال المهمة بنجاح يصبح أكثر صعوبة (Kilinc et al, 2021).

4.3.3. صياغة التعلم المعزز كعميلة قرار ماركوف (Markov Decision Process – MDP):

غالبا ما تصاغ مشكلة DRL كعميلة قرار ماركوف حيث يعرف وكيلا برمجيا يتفاعل مع البيئة ويتم نمذجة هذا التفاعل كسلسلة من المشاهدات والإجراءات في الوقت المناسب. في كل خطوة زمنية يتم وصف البيئة من خلال حالتها الحالية التي يلاحظها الوكيل وإلى جانب كل حالة يتلقى الوكيل أيضا مكافأة من البيئة التي تدل على أن هذه الحالة جيدة. كلما زادت المكافأة التي يحصل عليها الوكيل زادت خبرته الجيدة. إذن يكون الهدف من RL تعظيم المكافأة الإجمالية المتراكمة. وإذا كانت مهمة RL تفي بخاصيتي ماركوف وهما:

- احتمالية الحالة في الخطوة t + 1 بالنظر إلى الحالة في الخطوة t تساوي احتمالية الحالة عند t + a والتي تضم جميع الحالات من 1 حتى t، وهو ما توضحه العلاقة (4-5) التالية:

$$P\left(s_{t+1}|s_t\right) = P(s_{t+1}|s_1, \dots, s_2) \quad (4-5)$$

أي بالنظر إلى الحالة السابقة، فإن جميع الحالات التي تسبقها لا تهم أو بعبارة أخرى، نظرا للحاضر، فإننا لا نهتم بالماضي إذ تحتوي الحالة الحالية جميع المعلومات الضرورية من تاريخ الأحداث لذلك يمكن تجاهل ما يسبقها.

- كذلك إذا كانت البيئة يمكن مشاهدتها بالكامل من قبل الوكيل.

عندها يمكن صياغة المشكلة كعميلة قرار ماركوف لاتخاذ القرار MDPs: بافتراض أن MDPs محدودة فسيكون:

- **الحالة:** $s \in S$ حيث S مجموعة محددة من لحالات تصف حالة البيئة في لحظة زمنية t (step).
- **الإجراء:** $a \in A$ حيث A مجموعة محددة من الإجراءات.
- **دالة الانتقال أو احتمالات الانتقال:** $P_{s,a,s_0} = P(s_{t+1}|s_t, a_t)$ هي احتمال الانتقال من الحالة s إلى s_0 عند تطبيق الإجراء a. يمكن أن تكون مهمة RL إما مستندة إلى النموذج حيث يُعرف بـ MDP المستندة على النماذج ويسمى التخطيط. وتعرف فيه احتمالات الانتقال بشكل صريح وبالتالي يمكن للوكيل أمثلة سياسته دون التفاعل مع البيئة. أو MDP الخالية من النماذج. هذا هو السيناريو الأكثر شيوعا إذ البيئة غير معروفة في البداية ولا يوجد تعريف صريح لدالة الانتقال وبالتالي يحسن الوكيل سياسته من خلال التفاعل مع البيئة.
- **معامل الخصم:** $\gamma \in [0,1]$ كلما كانت قيمته أقرب للصفر قلت أهمية المكافأة المستقبلية.
- **المكافأة:** $R_{a,s,s_0} = r(s, a, s_0)$ أو $r_{t+1} = r\left(s_{t+1}, s_t, a_t\right)$، يتلقاها الوكيل عند الانتقال من حالة s إلى s_0 بتطبيق الإجراء a.

4.3.4. العناصر الأساسية لوكيل التعلم المعزز:

<u>السياسة Policy:</u> هي سلوك الوكيل وتمثل خرائط تُرسَم من الحالات إلى الإجراءات أو توزيع احتمالية الإجراء (مثل الاحتمالات لكل إجراء يجب اتباعه). السياسة π يمكن أن تكون:

- **حتمية Deterministic:** $a = \pi(s)$ تعطي ببساطة الإجراء a لحالة ما s.
- **عشوائية Stochastic:** $\pi(a|s) = P[A_t = a|\, S_t = s]$ تعيد احتمالية اختيار الإجراء a في حالة s.

كما قد تكون السياسة خضعت للتقييم **Evaluated** وتشير إلى التنبؤ Prediction وهذا يتعلق بتقييم سياسة معينة وتقدير مدى جودتها، أو سياسة خضعت للأمثلة **Optimized** وتشير إلى التحكم Control وهي تتعلق بإيجاد أفضل سياسة ممكنة وهي التي تؤدي إلى أعلى مكافأة.

بشكل عام، الهدف من مهمة التعلم المعزز هو إيجاد السياسة المثلى π^* التي تساعد الوكيل على اختيار الإجراءات التي ستزيد من مكافأته التراكمية. ليست كل أنواع RL قائمة على السياسة، هناك خوارزميات خالية من السياسة قائمة على القيمة، حيث يتم تقدير قيم الحالات وأزواج إجراء حالة من السياسة.

دالة القيمة Value function: تحدد دالة القيمة أو دالة قيمة الحالة للسياسة π على أنها المكافأة المستقبلية المتوقعة في حالة معينة، تخبر مدى "جودة" / "سوء" تلك الحالة. وتعطى بالعلاقة:

$$V^{\pi}(s) = \mathbb{E}_{\pi}[R_t|S_t = s] = \mathbb{E}[\sum_{k=0}^{\infty}\gamma^t r_{t+k+1}|S_t = s] \qquad (4\text{-}6)$$

حيث R_t هو إجمالي العائد المخصوم وr_{t+k+1} هي المكافأة الفورية. تدل هذه المعادلة على المكافأة الكلية التي يمكن الحصول عليها عند الوصول إلى الحالة s باتباع السياسة π.

دالة قيمة الإجراء Action Value function: هي مدى جودة إجراء يتخذ بالوجود في حالة ما باتباع سياسة π:

$$Q^{\pi}(s,a) = \mathbb{E}_{\pi}[R_t|S_t = s\,, A_t = a] = \mathbb{E}[\sum_{k=0}^{\infty}\gamma^t r_{t+k+1}|S_t = s\,, A_t = a] \qquad (4\text{-}7)$$

وهو العائد المتوقع عندما تكون الحالة s، ويتم اتخاذ الإجراء a بفعل السياسة π.

هناك أنواع متعددة من الخوارزميات لحساب $V\pi$ و$Q\pi$، بعضها يقدر $V\pi$ و$Q\pi$ بناء على الخبرة، والبعض الآخر يحسب القيم الحقيقية المتوقعة على افتراض أن نموذج MDP معروف.

نموذج البيئة: هو وصف لسلوك النظام، يستخدم للتنبؤ بما ستقوم به البيئة ويتكون من P وR حيث، تقوم P بتحديد احتمال الانتقال من حالة s في الخطوة الزمنية t إلى الحالة 's في الخطوة الزمنية t+1 عند اتخاذ الإجراء a في الحالة s، وR يستخدم للتنبؤ بمقدار المكافأة التي سيتم الحصول عليها في الخطوة t+1 عند التواجد في الحالة s واتخاذ الإجراء a في الخطوة الزمنية t. أي أن النموذج هو التنبؤ بالحالة والمكافأة القادمين عند التواجد في حالة s واتخاذ إجراء a. يمكّن من استنتاج كيفية تصرف البيئة، ويتنبأ بالحالة التالية، ومكافأة الحالة الحالية والإجراء.

غالبا ما يتم استخدام النماذج للتخطيط **Planning** في RL، ويعني ذلك أنه يستخدم لتعديل السياسة دون تفاعل مباشر مع البيئة، والتي تستلزم أنه من الممكن اتخاذ قرار بشأن الإجراءات المستقبلية التي يجب على الوكيل اتخاذها قبل أن يختبرها. تسمى طرق RL التي تستخدم النماذج للتخطيط الأساليب القائمة على النموذج. يشار إلى نظيرهم على أنه أساليب خالية من النماذج ويستخدم التجربة والخطأ لتعلم السياسة. إن وجود نموذج للبيئة يمنح الوكيل القدرة على التخطيط للمستقبل. ويجعل من الممكن للوكيل أن يشكل سياسة بناء على ما سيختبره في المستقبل. ومع ذلك لا يتم إعطاء الوكيل دائما نموذجا للبيئة، ولكن في كثير من الحالات يجب أن يتعلم **Learning** الوكيل النموذج من التفاعل مع البيئة (Zhang et al, 2021) (Danielsen et al, 2021).

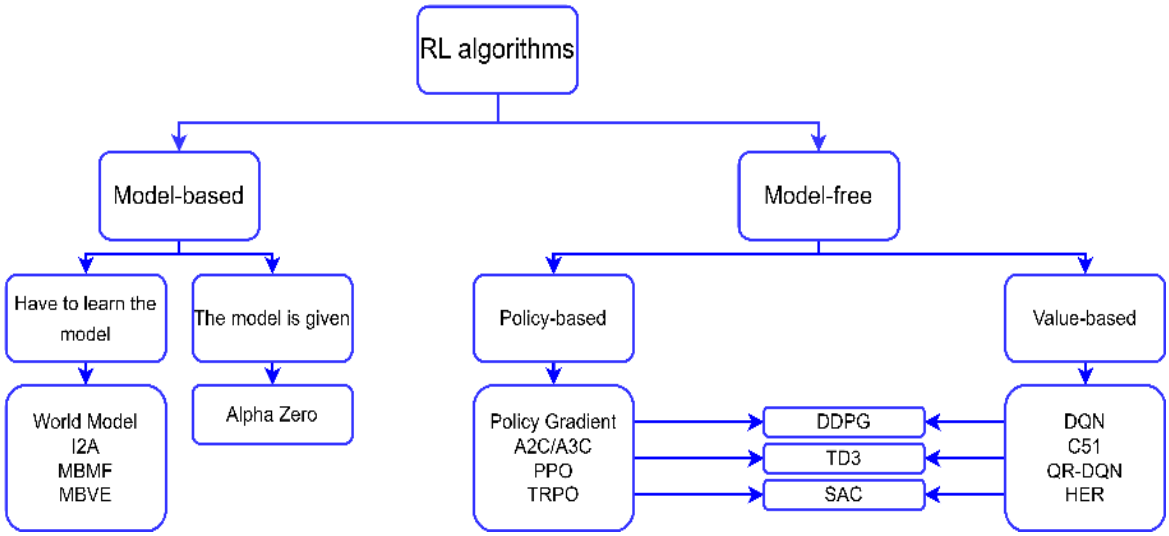
4.3.5. مناهج التعلم المعزز الكلاسيكية:

هي منهاج لتقدير توابع قيمة الحالة وتوابع قيمة الإجراء كما هو موضح في الشكل (6-4)، هناك ثلاث عائلات رئيسية من الخوارزميات المستخدمة في التعلم المعزز: البرمجة الديناميكية (Dynamic Programming - DP) ومونت كارلو (Monte Carlo - MC) والفرق الزمني (Temporal Difference - TD). تهدف طرق DP إلى إيجاد السياسة المثلى، لكنها تتطلب نموذج رياضي مثالي للنظام أو لوصف البيئة والقيود الحسابية للمهام تجعلها غير قابلة للتنفيذ في البيئات الكبيرة أو غير المعروفة. تكرار السياسة Policy Iteration وتكرار القيمة Value Iteration هما خوارزميتان شائعتان في البرمجة الديناميكية حيث يسعى تكرار السياسة إلى السياسة المثلى من خلال أمثلة السياسة المتكرر وتقييمها. ومع ذلك، نادرا ما يتم استخدامه بسبب تكلفته الحسابية الكبيرة. في المقابل، يحدد تكرار القيمة السياسة المثلى من خلال تحديث دوال القيمة وهي أسرع وأكثر كفاءة لأنها لا تقييم العديد من السياسات الكاملة بل تُحدث القيم مباشرة. رغم هذا فهي تُلزم وجود نموذج مثالي للنظام لاستخراج السياسة المثلى باستخدام دالة القيمة المثلى. لا تحتاج طرق MC لوجود نماذج مسبقة للبيئة وتعتمد على أخذ العينات من التجربة لتقدير متوسط المكافآت لمختلف السياسات من خلال أخذ عينات من تسلسلات الحالة والإجراءات والمكافآت بموجب السياسة. نظرا لأن الوكيل ليس لديه نموذج للنظام، فإنه يحدد وظائف القيمة لكل إجراء من خلال الاستكشاف ويستنتج السياسة المثلى. تحدد طرق MC وظائف القيمة والإجراء لأن وظائف القيمة وحدها غير كافية بدون نموذج للتبديل إلى حالات القيمة الأعظم. ومع ذلك، تتطلب خوارزميات MC الانتظار حتى نهاية الحلقة قبل أن تبدأ في تحديث دوال القيمة، مما يجعلها غير مناسبة للأنظمة التي تكون حلقاتها طويلة أو غير منتهية ومستمرة. ومع ذلك، فهي أكثر ملاءمة في البيئات التي تتسم بالعشوائية العالية أو غير النمطية، لأنها لا تعتمد على النموذج لكنها تتطلب عددا كبيرا من العينات للوصول إلى تقدير دقيق للقيم، مما قد يبطئ التعلم. عرّف Sutton عام 1988 فئة جديدة وهي الفرق الزمني TD تجمع بين كل من DP وMC باستخدام أفكار كلتا الطريقتين لتحديث وظائف القيمة بشكل تدريجي. يمكنها التعلم دون الحاجة إلى نموذج بيئي مثل مونت كارلو. لا تتطلب طرق TD الانتظار حتى نهاية الحلقة قبل أن يبدأ التعلم، وهو أمر غير محبب في الأنظمة الطويلة أو المستمرة كما ذكرنا، فتُحدث قيمها بعد كل خطوة. نظرا لأن الهدف الرئيسي للذكاء الاصطناعي هو تكرار السلوك البشري، فإن DP أو MC وحدهما غير كافيين، وغالبا ما تُستخدم طرق TD بدلا من ذلك. إذ تعد الأكثر كفاءة في التطبيقات العملية مثل الذكاء الاصطناعي في الألعاب والروبوتات (Sekkat et al, 2023). حيث اعتبرت تقدما كبيرا في مجال التنبؤ لأنها سمحت بتحديث التنبؤات بشكل تدريجي في بيئة ما وبشكل مباشر، باستخدام تقريب الوظيفة وإعادة التشغيل. تقريب الوظيفة وهو أسلوب تمثيل وظيفة معقدة أو غير معروفة باستخدام وظيفة معروفة، وغالبا ما تكون أبسط، ويعد أسلوب قوي في تعلم الآلة لتمثيل الحالة فهو يسمح للوكيل بضغط عدد كبير أو حتى لا نهائي من الحالات في مجموعة محدودة من المحددات، في حين أن ضغط الحالة بهذه الطريقة يمكن أن يؤدي إلى فقدان المعلومات، إلا

أنه غالبا ما يكون الطريقة الوحيدة لجعل المشكلات قابلة للحل وفعّالة حسابيا. علاوة على ذلك، يتمتع تقريب الوظيفة بميزة إنشاء تعميم عبر الحالات، مما يمكن أن يسرع التعلم. يتناقض استخدام تقريب الوظيفة مع الأساليب الجدولية حيث يتم تمثيل كل حالة بشكل فريد في شكل ما من أشكال جدول البحث. لا تُستخدم مثل هذه الأساليب الجدولية في هذه الأطروحة، ولكن تمت مناقشتها بشكل مفصل في (Sutton and Barto, 1998). في حين أن مصطلح إعادة التشغيل bootstrapping له معان عديدة، فإنه يُستخدم هنا كما هو في البرمجة الديناميكية، أي أننا نستفيد من التنبؤات الموجودة لتحديث تنبؤات أخرى. وهذا يعني فعليا أنه يمكن إجراء تحديثات للتنبؤات قبل رؤية النتيجة المتوقعة يعد طريقة فعّالة لتحديث التنبؤات ومع ذلك يمكن أن يكون له تحيز كبير (أخطاء من تقديرات غير دقيقة). وعلى الطرف الآخر يتم إجراء تحديثات MC فقط في نهاية الحلقة، بمجرد الوصول إلى الحالة النهائية. يمكن أن يكون لخوارزميات MC تباين كبير في المجالات العشوائية فهي حساسة للاختلافات في المكافآت (Sherstan, 2015).

4.3.6. خوارزميات التعلم المعزز:

من الصعب تقديم قائمة شاملة ومفصلة لجميع خوارزميات RL وخاصة تلك المطبقة على الروبوتات. لذا تُذكر الفروع الرئيسية للخوارزميات هنا والتي تشمل الخوارزميات القائمة على النموذج والخالية من النموذج وكذلك الخوارزميات القائمة على السياسة والقائمة على القيمة. يمثل الشكل (6-4) قائمة كاملة بالخوارزميات. في RL القائم على النموذج، تُستخدم العينات لتعلم نموذج ديناميكي للبيئة، والذي يُستخدم بدوره في خوارزمية التخطيط أو التحكم الأمثل لإنتاج سياسة أو تسلسل عناصر التحكم. في RL الخالي من النماذج، لا يتم نمذجة الديناميكيات بشكل صريح، ولكن بدلا من ذلك يتم تعلم السياسة المثلى أو دالة القيمة مباشرة من خلال التفاعل مع البيئة.



الشكل 6-4 خوارزميات التعلم المعزز العميق (Sekkat et al, 2023).

من الشكل (6-4) نرى أنه يمكن تصنيف خوارزميات RL المعتمدة على النموذج إلى مجموعتين اعتمادا على ما إذا كان الوكيل لديه معرفة بنموذج البيئة أو يتعلمه. إذا كان لديه النموذج، فيمكنه التنبؤ بأدائه عند اتخاذ إجراء

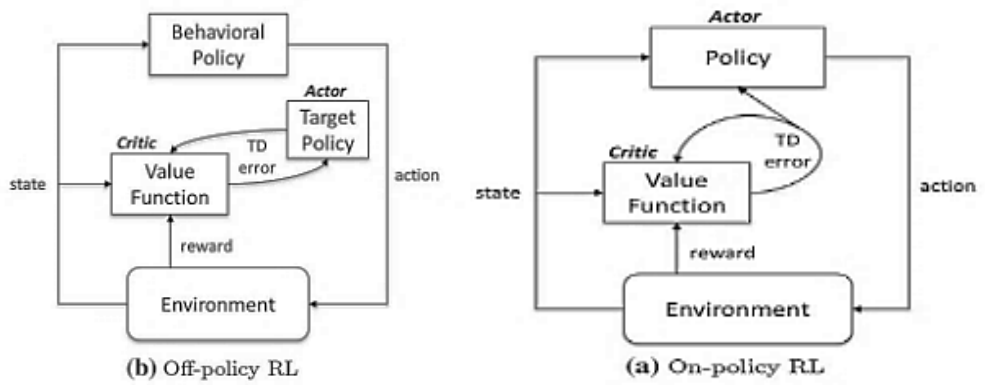
معين، مما يؤدي إلى أمثلة كفاءة العينة مقارنة بالطرق الخالية من النموذج. ومع ذلك، فإن تعلم النموذج يمكن أن يؤدي إلى تحيز، مما يؤدي إلى سلوك دون المستوى الأمثل في البيئة الحقيقية. وعلى النقيض من ذلك، تعتمد تقنيات الخوارزميات الخالية من النموذج على إشارات المكافأة وتتعلم توابع القيمة فقط من تفاعلات الوكيل مع البيئة. إنها أسهل في التنفيذ والتعديل للمحددات الفائقة، مما يجعلها أكثر شيوعا من الأساليب المعتمدة على النموذج. كذلك يمكن تصنيف سياسات التعلم المعزز الخالية من النموذج إلى سياسات تعتمد على القيمة في الخوارزميات التي تعتمد على القيمة، وسياسات تعتمد على السياسة في الخوارزميات التي تعتمد على السياسة. تركز الأساليب القائمة على السياسات على تحديد السلوك الأمثل بشكل مباشر، وتقدر الأساليب القائمة على القيم قيمة الإجراءات المختلفة في حالة معينة. ففي الخوارزميات المعتمدة على القيمة، يتم تقدير دالة الإجراء والقيمة بالإشارة إلى القيمة المثلى $Q^*(s, a)$. يتم تحقيق ذلك عادة من خلال التعلم خارج السياسة OFF-Policy. وعلى العكس من ذلك، تحدد الأساليب المعتمدة على السياسة الإجراء الأمثل الذي يجب اتخاذه في حالة (حالات) معينة لتعظيم المكافأة غالبا من خلال التعلم المباشر داخل السياسة ON-Policy. إن التقنيات القائمة على السياسة أكثر موثوقية وثباتا من أساليب تعلم Q (القيمة) والتي تقدر Q بشكل غير مباشر بناء على دالة موضوعية ومع ذلك، قد تفشل الأساليب القائمة على السياسة بسبب عوامل مختلفة (Sekkat et al, 2023) والجمع بين النهجين ينتج نهج الممثل الناقد.

4.3.6.1. مفهوم ON-Policy وOFF-Policy:

<u>التعلم المباشر (داخل السياسة) On Line Policy:</u> عندما يعتمد RL على التفاعل مع البيئة، يتم اختيار الإجراءات مع ما يسمى بالسياسة السلوكية، المشار إليها μ. هذه السياسة ليست بالضرورة هي السياسة التي نريد أمثلتها أو تقييمها (المشار إليها π). إذا كانت $\pi = \mu$، فإن السياسة المراد تقييمها هي السياسة المستخدمة للتفاعل مع البيئة. يتعلم الوكيل أثناء العمل، وبالتالي هي خوارزمية مرتبطة بالسياسة، ويتم تعلم أفضل سياسة استنادا إلى السياسة التي تم تعلمها مسبقا (Kepes, 2017) (Gallouedec et al, 2020).

<u>التعلم غير المباشر (خارج السياسة) OFF line Policy:</u> عندما يستخدم الوكيل سياسة مختلفة لأخذ عينات من الحلقات، نكون في حالة خوارزمية تعلم خارج السياسة. من المزايا الرئيسية للتعلم خارج السياسة أنه يمكن استخدام كل تجربة: الخبرة السابقة بالإضافة إلى خبرة أي وكيل آخر. يتعلم أفضل سياسة بغض النظر عن سياسة السلوك، وهي أنسب لتطبيقات الروبوتات (الإجراءات المستمر). وتمكن من استخدام كافة البيانات التي تم جمعها في اللحظات السابقة حتى اللحظة الحالية لتدريب السياسة، وبالتالي، من الممكن توسيع نطاقها لتشمل مليارات العينات. لكن في معظم معايير RL، تجمع هذه الأساليب بيانات جديدة أثناء إجراء التدريب، ويمكن تدريبها دون جمع أي بيانات؛ على غرار التعلم الموجه (Ibarz et al, 2021) (Kepes, 2017) (Gallouedec et al, 2020).

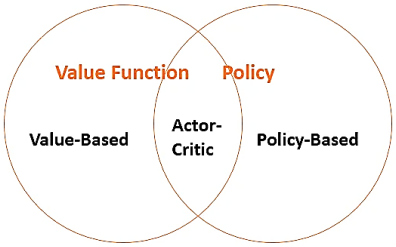
يوضح الشكل (7-4) التالي الفرق بين التصنيفين.



الشكل 7-4 الفرق بين التصنيفين (Shahid et al, 2022).

4.3.6.2. مفهوم أسلوب الفاعل – الناقد Actor -Critic:

تقسم خوارزميات RL الخالية من النماذج إلى ثلاث أنواع Value Based، Policy Based، Actor – Critic كما يوضح الشكل (4-8)، ولفهم نهج الفاعل الناقد المستخدم في هذا البحث لا بد من إلقاء نظرة على المفهومين الآخرين، لذا تقدم مقارنة للأساليب الثلاث (Yu et al, 2022):



الشكل 8-4 موقع أسلوب الفاعل – الناقد بين الأساليب الأخرى
(https://medium.com/data-science/birds-eye-view-of-reinforcement-learning-algorithms-landscape-2aba7840211c).

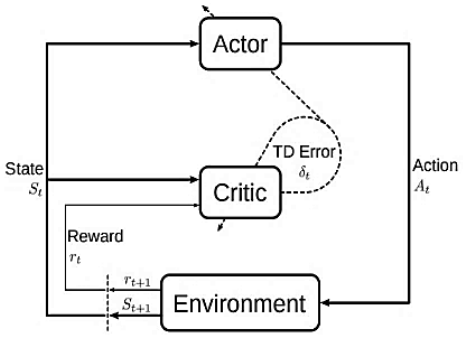
في النهج القائم على القيمة يتم تعلم دالة القيمة وتحديد السياسة من خلال استراتيجية الجشع (Greedy Strategy) أو استراتيجية تعلم DQN وهي طريقة نموذجية قائمة على القيمة. بالرغم من أنها حققت أداء ممتازا في مشاكل مساحة عمل منفصلة مثل ألعاب الفيديو وحتى هزم اللاعبين البشر إلا أنها غير قادرة على التعامل مع مشاكل ذات مساحة العمل المستمر الموجودة في العديد من مهام الحياة اليومية كالتلاعب الماهر. تختلف الخوارزميات القائمة على السياسة عن النهج القائم على القيمة بأن السياسة تتحسن مباشرة ومن خوارزمياتها REINFORCE PPO TRPO كما تحتاج الخوارزميات المعتمدة على السياسة إلى العديد من العينات إذ أنه لتحديث السياسة الحالية يجب أخذ عينات من التدريب وضمان تقارب الخوارزمية في كل مرة يتم بها تحديث السياسة. أغلب الخوارزميات الحالية لديها كفاءة أخذ عينات منخفضة وتحتاج قوة حسابية لدعمها لذا فإن الجمع بين الطريقتين السابقتين يدعى نهج الفاعل الناقد يمكنه أن يحل هذه المشكلة إلى حد ما، خوارزمية الفاعل الناقد في الغالب تكون خارج السياسة ويمكنها حل مشكلة أخذ العينات من خلال نهج إعادة التجربة. ومع ذلك فإن اقتران تحديث السياسة وتقييم القيمة يؤدي إلى عدم استقرار للخوارزمية وخاصة حساسية المحددات الفائقة. كما يكون من الصعب جدا ضبط المحددات وتواجه هذه الخوارزميات صعوبة في النشر عندما يتم ترقيتها إلى مجال التطبيق (Yu et al, 2022). فعلى النقيض

من الطرق القائمة على القيمة والسياسة كفئتين أساسيتين تتضمن طرق الفاعل والناقد خوارزميات تستخدم كلا من السياسة المُعلمة أي الفاعل الذي يتعلم السياسة أي كيفية اختيار الأفعال ودالة القيمة أي الناقد الذي يتعلم دالة القيمة أي تقييم الإجراء الذي اختاره الفاعل. ويتم تحقيق ذلك باستخدام شبكات منفصلة حيث يمكن للفاعل والناقد في بعض الأحيان مشاركة بعض المحددات المشتركة. يسمح هذا المزيج لخوارزميات الفاعل والناقد أن تتمتع في وقت واحد بمزايا كلا النهجين مثل كفاءة العينة ومساحة العمل المستمر. لذلك، جعلت هذه الخصائص طرق الفاعل والناقد شائعة للتلاعب الآلي بالأشياء مع تحقيق أداء متطور بين طرق RL الأخرى في هذا المجال. على غرار الطرق القائمة على السياسة، تتعلم شبكة الفاعل احتمالية اختيار إجراء a معين في حالة معينة s بناء على محددات وأوزان شبكته θ أي $\pi(a|s,\theta)$. تستخدم شبكة فاعل خطأ نهج الفروق الزمنية TD error لحساب مدى جودة الإجراء الذي تم اختياره، وهو مقدار الخطأ بين القيمة الحقيقة الناتجة عن السياسة وبين القيمة المتوقعة والذي يستخدم لتحديث التقدير وجعل النموذج يتعلم، ويعطى هذا الخطأ في دالة قيمة الإجراء بالعلاقة:

$$\delta_t = r_t + \gamma \underbrace{max}_{\grave{a}} Q^{\pi}(s_t, \grave{a}) - Q^{\pi}(s_t, a) \qquad (4\text{–}8)$$

- r_t مكافأة الوكيل في الخطوة الحالية.
- γ معامل خصم ويحدد مدى أهمية المكافآت المستقبلية.
- $\underbrace{max}_{\grave{a}} Q^{\pi}(s_t, \grave{a})$ القيمة المتوقعة في الحالة التالية.
- $Q^{\pi}(s_t, a)$ القيمة الحالية المقدرة للإجراء في الحالة الحالية.

ثم تحدث شبكة الناقد دالة قيمة الإجراء $Q(s,a)$ عن طريق تقليل هذا الخطأ. يجب أن تكون القيم المقدرة والفعلية متطابقتين ولكن في الواقع لا تكون متطابقة وبالتالي يظهر الفرق (الخطأ) وتستخدم هذا الخطأ لانتقاد الفاعل على أساس مدى جودة الإجراء المحدد. تم توضيح هذه العملية في الشكل (4–9). ومع ذلك، يُقال إن الترابط المتبادل لتوزيع مخرجات كل منهما يمكن أن يؤدي إلى عدم الاستقرار أثناء التعلم ويجعل من الصعب ضبطها. وعلى الرغم من ذلك، يتم استخدام خوارزميات التعلم المعزز الخالية من النماذج التي تعتمد على الفاعل والناقد في الكثير من الأبحاث (Orsula, 2021) ومنهم هذ البحث.



الشكل 4–9 منهجية الفاعل – الناقد حيث يتم استخدام TD error لتعديل سياسة π(a|s,θ) ودالة قيمة Q(s|a) للفاعل (Orsula, 2021)(Sutton and Barto, 2018).

4.3.7. مفهوم الاستكشاف والاستغلال Exploitation and Exploration:

الهدف في خوارزمية RL هو إيجاد سياسة مثالية تحقق المهمة ولكن التحدي هو كيفية تعظيم مجموع المكافآت لذا فإن الاستكشاف والاستغلال يشكلان معضلة لنظام RL. يتعلق الاستغلال بالتزام الوكيل بما فهمه من البيئة حتى الآن وقبوله كسياسة جيدة ولكن الخطر هو تفويت فرص أخرى غير السياسة التي تعلمها والتي قد تؤدي إلى الحصول على المزيد من المكافآت. من ناحية أخرى، يتعلق الاستكشاف بجعل الوكيل يبحث في بالبيئة على أمل استكشاف المزيد من المكافآت ولكن بالطبع الخطر هو إضاعة الوقت والحصول على ردود فعل سلبية. بشكل عام، يتوجب على الوكيل استكشاف بيئته من أجل تعلم أفضل ثم استغلال معرفته المتشكلة عن الاستكشاف. للموازنة بين الاستكشاف والاستغلال (متى يجب على الوكيل استكشاف البيئة أكثر ومتى عليه استغلال معرفته وتقليل الاستكشاف وجمع المعلومات) طوّرت العديد من الخوارزميات معظمها يعتمد على نهج الجشع المشترك حيث يساعد عامل جاما وهو رقم بين 0 و1 الوكيل على اكتشاف بيئته في البداية ولكن بمرور الوقت عندما يتقدم التعلم فإنه يقلل من احتمالية الاستكشاف ويتحرك أكثر نحو الاستغلال. حيث يختار الوكيل إما الإجراء ذو أعلى قيمة (الاستغلال) أو إجراء عشوائي (الاستكشاف) ومن هذه الطرق: ε-Greedy Method و Softmax Action Strategy وOptimistic Initial Values (Korber et al, 2021) (Mellatagshi et al, 2021).

4.3.8. أساليب أمثلة السياسة Policy Optimization:

4.3.8.1. خوارزمية تدرج السياسة Policy Gradient:

تقوم بحساب قيمة تقديرية توضح كيف يجب أن تعدل السياسة لأمثلة الأداء، ثم تستخدم هذه القيمة لتحديث النموذج تدريجيا نحو الأفضل، من خلال إدخالها في خوارزمية صعود التدرج العشوائي (Stochastic Gradient Ascent – SGA) ويستخدم الصعود بسبب الرغبة بتعظيم الأداء، وتستخدم العشوائية لتدل على أن الخوارزمية تستخدم عينات عشوائية من البيئة، وليس كل البيانات دفعة واحدة. يعطى مقدر التدرج الأكثر شيوعا بالعلاقة:

$$L^{PG} = \hat{g} = \hat{E}_t\left[\nabla_\theta \, log \, \pi_\theta(\alpha_t \,|s_t)\hat{A}_t\right] \quad (4-9)$$

حيث π_θ هي سياسة عشوائية و$\hat{A}_t$ هي مقدر دالة الميزة Advantage Estimator عند خطوة زمنية t. والميزة (Advantage) تعبّر عن مدى جودة الإجراء a في الحالة s مقارنة بما هو متوقع من تلك الحالة. والصيغة العامة لها هي: $\hat{A}_{(s,\,a)} = Q_{(s,\,a)} - V_{(s)}$، وتكون Q (s, a) القيمة المتوقعة لأخذ الإجراء a في الحالة s. و$V_{(s)}$ القيمة المتوقعة لجميع الأفعال في الحالة s أي المتوسط. والفرق بينهما يعطي الميزة لهذا الإجراء. عادة تكون Q وv غير معلومين بدقة لذا نحتاج لمقدر للميزة (Estimator). تشير $\hat{E}_t$ إلى المتوسط التجريبي على دفعة محدودة من البيانات، في خوارزمية تتناوب بين أخذ العينات والأمثلة، أي هي التوقع أو المتوسط الإحصائي عبر عينات كثيرة. الخلاصة: تُستخدم دالة هدف قائمة على تقدير الميزة، تمثل كم كان الإجراء جيدا، وتشجع السياسة على تكراره

تعرف بدالة هدف $L^{PG}(\Theta)$ تعني Policy Gradient Loss، وكلما كانت السياسة تعطي احتمال أعلى للإجراء يعني $(\hat{A}_t)$ أكبر، كلما زادت قيمة الهدف. يكون مشتقها هو مقدر تدرج السياسة (policy gradient) أي يتم اشتقاقها آليا للحصول على المقدر $\hat{g}$ ويستخدم كتدرج سياسة لتحديث المعاملات Θ لكن يجب الانتباه لألا تكرر التحديثات عدة مرات على نفس البيانات، لأنه يسبب تغييرات كبيرة وغير مستقرة (Schulman et al, 2017).

4.3.8.2. خوارزمية منطقة الثقة لأمثلة السياسة - Trust Region Policy Optimization :TRPO

يتم فها أمثلة سياسات RL بطريقة آمنة تحافظ على استقرار الأداء وتجنب التغيرات الكبيرة في كل تحديث. فيها يتم تعظيم دالة الهدف البديل surrogate objective وفقا لقيود على حجم تحديث السياسة. وبالتالي التأكد من أن التحديث لا يبتعد كثيرا عن السياسة السابقة. يتم ذلك من خلال تعظيم الهدف البديل التالي:

$$\underbrace{maximize}_{\theta} \hat{E}_t \left[\frac{\pi_\theta(\alpha_t|s_t)}{\pi_{\theta\ old}(\alpha_t|s_t)} \hat{A}_t \right] \qquad (4-10)$$

$\pi_\theta(\alpha_t|s_t)$: السياسة الحالية بعد التحديث. $\pi_{\theta\ old}(\alpha_t|s_t)$: السياسة السابقة قبل التحديث. $\hat{A}_t$: الميزة التي تقيس مقدار جودة الإجراء في هذه الحالة. يعبر الكسر عن نسبة احتمالات السياسة الجديدة إلى القديمة. ويقيس الهدف البديل كم تحسنت السياسة في الأداء بناء على المعلومات السابقة.

القيد: يستخدم القيد لقمع التغيرات الكبيرة في السياسة الجديدة نسبة للسياسة القديمة، لأن التحديثات الكبيرة قد تسبب انهيار في الأداء أو سلوك غير مستقر، وبالتالي ضمان أن كل تحديث يبقى ضمن منطقة الثقة. ويتم قياس قيد Kullback–Leibler (KL) Divergence وهو مقياس الفرق بين السياسات السابقة والجديدة، بالعلاقة:

$$\text{Subject to } \hat{E}_t[KL[\pi_{\theta\ old}(.|s_t), \pi_\theta(.|s_t)]]\hat{A}_t \leq \delta \qquad (4-11)$$

حيث $\pi_{\theta\ old}$ يمثل متجه محددات السياسة قبل التحديث. و δ حد يحدد منطقة الثقة. يصبح حل هذه المشكلة بكفاءة باستخدام خوارزمية التدرج المترافق conjugate gradient ممكنا، بعد إجراء تقريب خطي للهدف وتقريب تربيعي للقيد. كما تقترح خوارزمية TRPO استخدام عقوبة بدلا عن القيد عند الابتعاد الكبير عن السياسة القديمة، يعرف معامل العقوبة بـ β.

$$\underbrace{maximize}_{\theta} \hat{E}_t \left[\frac{\pi_\theta(\alpha_t|s_t)}{\pi_{\theta\ old}(\alpha_t|s_t)} \hat{A}_t - \beta KL[\pi_{\theta\ old}(.|s_t), \pi_\theta(.|s_t)] \right] \qquad (4-12)$$

تفضل خوارزمية TRPO القيد الثابت بدل العقوبة لأنه أكثر استقرار وموثوقية، ويصعب اختيار قيمة لـ β تعمل بشكل جيد عبر مشكلات مختلفة أو داخل مشكلة واحدة، حيث تتغير الخصائص على مدار التعلم. لذلك يكون

هناك حاجة لتعديلات إضافية عند استخدام العقوبة بدلا عن القيد للوصول إلى نفس أمثلة الأداء المضمون، مفهومي القيد والعقوبة ستشرح لاحقا بشكل أوسع.

4.3.8.3. دالة الهدف البديل المقطوعة Clipped Surrogate Objective:

لتكن $r_t(\theta)$ النسبة الاحتمالية $r_t(\theta) = \frac{\pi_\theta(\alpha_t|s_t)}{\pi_{\theta\ old}(\alpha_t|s_t)}$، إذا $r_t(\theta_{old}) = 1$. تقوم TRPO بتعظيم الهدف البديل:

$$L^{CPI}(\theta) = \hat{E}_t\left[\frac{\pi_\theta(\alpha_t|s_t)}{\pi_{\theta\ old}(\alpha_t|s_t)}\hat{A}_t\right] = \hat{E}_t[\ r_t(\theta)\ \hat{A}_t] \qquad (4\text{-}13)$$

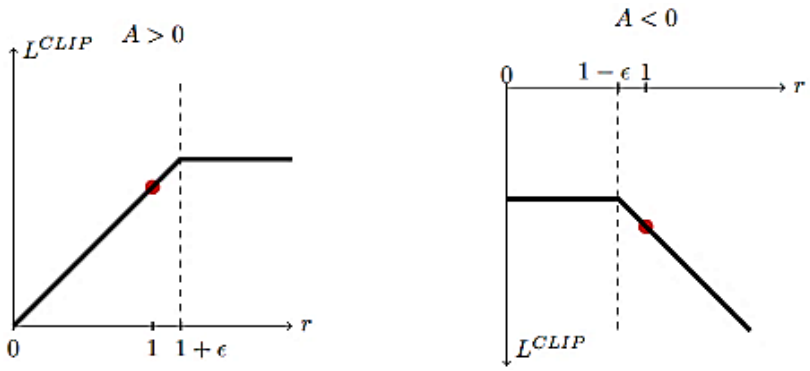
تشير النسبة الاحتمالية CPI إلى Conservative Policy Iteration، أي "تكرار السياسات المحافظ"، وهو نهج تم اقتراحه سابقا يتم فيه أمثلة السياسة تدريجيا بشكل محافظ، أي بدون تغييرات كبيرة. لو تم تعظيم الدالة الهدف L^{CPI} وهي تعبر عن فائدة أو جودة السياسة الجديدة بدون أي قيود، فقد يؤدي ذلك إلى تحديث كبير جدا للسياسة. هذا غير مرغوب فيه، لأن التغييرات الكبيرة يمكن أن تؤدي إلى نتائج غير مستقرة أو سيئة في بيئة RL. لهذا السبب، تم تعديل دالة الهدف بحيث تعاقب التغيرات الكبيرة في السياسة وبالتحديد تضيف نوعا من العقوبة إذا ابتعدت النسبة $r_t(\theta)$ عن القيمة 1 وهي تمثل نسبة احتمالات السياسة الجديدة إلى القديمة أما إذا كانت قريبة من 1 فهذا يعني أن التغيير في السياسة صغير وآمن. وبالتالي تكرار السياسات المحافظ CPI قد يؤدي لتغيرات كبيرة جدا في السياسة لذا يجب إضافة قيود أو عقوبات لتجنب التحديثات الكبيرة والخطيرة وذلك بالحفاظ على $r_t(\theta) \approx 1$.

لذا اقترح John Schulman وزملاءه عام 2017، تعديل دالة الهدف L^{CPI} وفق التالي:

$$L^{CLIP}(\theta) = \hat{\mathrm{E}}_t\left[min\left(r_t(\theta)\hat{\mathrm{A}}_t,\ \ clip(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{\mathrm{A}}_t\right)\right] \qquad (4\text{-}14)$$

حيث ϵ هو معامل فائق، (معامل قطع يحدد الحد الأعلى والأدنى للتغير) يتم فيه الاختيار بين القيمة الأصغر بين الهدف غير المقطوع L^{CPI} ، والهدف المقطوع $clip(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{\mathrm{A}}_t$ ويقوم بقص النسبة الاحتمالية مما يمنع الحافز لتحركها خارج المجال $[1+\epsilon, 1-\epsilon]$ تكون $L^{CLIP} = L^{CPI}$ عندما r = 1 وتصبح مختلفة كلما ابتعدت السياسة الجديدة عن القديمة، يعرض الشكل (4-10) دورة واحدة (خطوة زمنية واحدة) من L^{CLIP}، يمثل المحور الأفقي قيم النسبة الاحتمالية r، والعمودي قيم دالة الهدف البديلة L^{CLIP}، النقطة الحمراء هي الوضع الابتدائي r = 1 حيث السياسة الجديدة = القديمة. ويتم التمييز بين حالتين:

1) عندما تكون الميزة موجبة (التحديث مفيد) يتم التحديث بشرط عدم تجاوز الحد الأعلى وهو $1+\epsilon$.
2) عندما تكون الميزة سالبة (التحديث ضار) يتم قمعه إذا قلّ عن $1-\epsilon$.



الشكل 10–4 خطوة زمنية واحدة للدالة L^{CLIP} كتابع للنسبة r الاحتمالية، يوضح منحني اليمين المخطط مع ميزة إيجابية، واليسار مع ميزة سالبة، وتظهر النقطة الحمراء نقطة بداية الأمثلة r=1 (Schulman et al, 2017).

الجدول 1–4 جدول لستة حالات مختلفة لدالة الهدف المقطوعة (Bick et al, 2021).

$r_t(\theta) > 0$	A_t	Return Value of min	Objective is Clipped	Sign of Objective	Gradient
$r_t(\theta) \in [1-\varepsilon, 1+\varepsilon]$	+	$r_t(\theta)A_t$	No	+	√
$r_t(\theta) \in [1-\varepsilon, 1+\varepsilon]$	-	$r_t(\theta)A_t$	No	-	√
$r_t(\theta) < 1-\varepsilon$	+	$r_t(\theta)A_t$	No	+	√
$r_t(\theta) < 1-\varepsilon$	-	$(1-\varepsilon)A_t$	Yes	-	0
$r_t(\theta) > 1+\varepsilon$	+	$(1+\varepsilon)A_t$	yes	+	0
$r_t(\theta) > 1+\varepsilon$	-	$r_t(\theta)A_t$	No	-	√

كما يبين الشكل (10–4) والجدول (1–4)، إن أخذ الحد الأدنى من الهدف المقطوع وغير المقطوع يعني سيتم اختيار إما الهدف المقطوع أو غير المقطوع بناء على النسبة وحالة الميزة. الحالة 1 و2: لا يتم تطبيق القطع نظرا لأن النسبة تقع داخل النطاق $[1-\varepsilon, 1+\varepsilon]$ في الحالة 1، يوجد ميزة إيجابية: الإجراء أفضل من متوسط جميع الإجراءات في تلك الحالة. ولذلك، ينبغي تشجيع السياسة الحالية على زيادة احتمال اتخاذ هذا الإجراء في تلك الحالة. في الحالة 2، الميزة سلبية: الإجراء أسوأ من متوسط جميع الإجراءات في تلك الحالة. ولذلك، ينبغي قمع السياسة الحالية عن اتخاذ هذا الإجراء في تلك الحالة. الحالة 3 و4: نسبة الاحتمال أقل من$[1-\varepsilon]$، فإن احتمال اتخاذ هذا الإجراء في تلك الحالة أقل بكثير من السياسة القديمة. إذا كان تقدير الميزة موجبا (A>0)، كما هو الحال في الموقف 3، فيرغب بزيادة احتمال اتخاذ هذا الإجراء في تلك الحالة. ولكن إذا كان تقدير الميزة سالبا، كما هو الحال في الحالة 4، فيرغب بتقليل احتمالية اتخاذ هذا الإجراء في تلك الحالة بشكل أكبر. ولذلك، فإن التدرج = 0، ويدل الخط المسطح، أنه لا يقام بتحديث الأوزان. الحالة 5 و6: النسبة أعلى من النطاق، إذا كانت نسبة الاحتمال أعلى من $[1+\varepsilon]$، فإن احتمال اتخاذ هذا الإجراء في تلك الحالة في السياسة الحالية أعلى بكثير مما كان عليه في السياسة السابقة. إذا كانت الميزة إيجابية، كما في الحالة 5، فلا يجب أن يكون هنالك جشع كبير. يوجد بالفعل

احتمالية أكبر لاتخاذ هذا الإجراء في تلك الحالة مقارنة بالسياسة السابقة. ولذلك، فإن التدرج = 0 بالتالي لا يتم تحديث الأوزان. إذا كانت الميزة سلبية، كما في الحالة 6، فيجب تقليل احتمالية اتخاذ هذا الإجراء في تلك الحالة (Bick et al, 2021). إذا يتم تحديث السياسة فقط عندما تكون النسبة ضمن حدود الهدف، أو النسبة خارج المجال ولكن الميزة تقرب أكثر للمجال، كأن يكون تحت النسبة ولكن الميزة موجبة، أو أن يكون فوق النسبة ولكن الميزة سالبة.

تعتمد خوارزمية أمثلة السياسة القريبة (Proximal Policy Optimization – PPO) المستخدمة في هذا البحث، على نهج الهدف المقصوص هذا، كما سنشرح لاحقا.

4.3.8.4. معامل العقوبة التكيفي Kullback–Leibler - KL:

نهج آخر يمكن استخدامه كبديل لدالة الهدف المقصوص أو بالإضافة إليه في PPO هو استخدام عقوبة KL على مقدار التباين بين السياسة الجديدة والقديمة. وتعديل معامل العقوبة بحيث يحقق تباين KL قيمة مستهدفة معينة والتي يرمز لها بـ d_{targ} عند كل تحديث للسياسة إذا كان KL أقل من d_{targ} تقلل العقوبة وإلا فتزداد. رغم أن هذه الإضافة أعطت لدى Schulman أداء أسوأ من دالة الهدف السابقة ومع ذلك يلقى الضوء عليها لأنها تعد معيار مهم وأساسي. مقياسKL Divergence، يقيس مدى تغير السياسة الجديدة عن القديمة، إذا كانت هذه القيمة صغيرة نسبيا يعني أن السياسة لم تتغير كثيرا والتحديث آمن. إذا كانت كبيرة يعني أن التغير كبير وقد يكون هذا خطرا على الاستقرار. في أبسط تطبيق يتم استخدم عدة حلقات من انحدار التدرج العشوائي على دفعات صغيرة minibatch (Stochastic Gradient Descent - SGD)، لأمثلة الهدف الذي يحتوي على عقوبة تباعد، وتعطى دالة الهدف مع العقوبة بالعلاقة:

$$L^{KLPEN}(\theta) = \hat{\mathrm{E}}_t\left[\frac{\pi_\theta(\alpha_t|s_t)}{\pi_{\theta\ old}(\alpha_t|s_t)}\hat{A}_t - \beta KL\left[\pi_{\theta\ old}(.|s_t), \pi_\theta(.|s_t)\right]\right] \qquad (4\text{-}15)$$

ويحسب $d = \hat{E}_t[KL[\pi_{\theta\ old}(.|s_t), \pi_\theta(.|s_t)]]$:

- إذا كانت $d<d_{targ}/1.5$ يكون $\beta \leftarrow \beta/2$
- إذا كانت $d<d_{targ}/1.5$ يكون $\beta \leftarrow \beta \times 2$

يتم استخدام المعامل β الجديد لتحديث السياسة الجديدة. مع هذا الأسلوب قد يلاحظ أحيانا تحديث للسياسة عندما يكون تباعد KL كبيرا جدا عن القيمة الهدف، ومع ذلك، فإن هذه الحالات نادرة، ومعامل β يتكيف بسرعة. القيمتان 1.5 و2 المستخدمتان تم تعينهم من خلال دراسة Schulman وزملائه عام 2017 ولا تؤثران على الأداء، كما أن β هي معامل فائق ليست بأهمية كبيرة عمليا إذ تقوم الخوارزمية بتعديله بسرعة. إذا الفكرة هي تكيف الخوارزمية تلقائيا للحفاظ على تباعد KL حول قيمة مستهدفة (Schulman et all, 2017).

4.3.9. خوارزمية أمثلة السياسة القريبة (Proximal Policy Optimization – PPO):

تعريف بخوارزمية PPO: تعتبر خوارزمية PPO (model-free online policy gradient reinforcement learning algorithm)، وقبل وجودها كانت خوارزمية TRPO فعالة ولكنها معقدة جدا في التنفيذ، لذلك سعى فريق من OpenAI وبقيادة John Schulman عام 2017 لتحسين الوضع حينها من خلال تقديم خوارزمية تحقق كفاءة البيانات والأداء الموثوق به لخوارزمية TRPO لكن تتجاوز مشاكلها، فتكون أسهل في التنفيذ. وتستخدم أمثلة من الدرجة الأولى فقط. فجاءت نتيجة لذلك PPO كبديل أبسط وأكثر كفاءة، وتحسّن الاستقرار أثناء التدريب بدون الحاجة لحسابات معقدة مثل TRPO. من مميزاتها: سهولة التنفيذ، وكفاءة استخدام العينة، تعمل بشكل جيد ضمن الكثير من البيئات، سهولة ضبط المحددات، تحقق توازنا بين الاستكشاف والاستغلال، وتفوقت على طرق تدرج السياسة ON-policy الأخرى وحققت توازن إيجابي بين تعقيد العينة والبساطة ووقت التنفيذ الفعلي. كما تعزز هذه الخوارزمية استقرار التدريب وكفاءته، خصوصا في المهام ذات الطبيعة الديناميكية والمعقدة. إذ تعد مثالية لتطبيقات الروبوتات التي تتطلب تحكما دقيقا وسلوكا معقدا، كما تضمن تعلم متسق وموثوق عن طريق استخدام دالة الهدف البديلة وقيود السياسات لتقليل التغيرات المتطرفة (Schulman et al, 2017).

مبدأ عمل خوارزمية PPO: تعتمد خوارزمية PPO على نهج الهدف المقصوص ودالة القيمة لتقدير ميزة اتخاذ إجراء في كل حالة. وتتناوب هذه الخوارزمية بين أخذ عينات من البيانات المجمعة من خلال التفاعل مع البيئة، وأمثلة دالة الهدف باستخدام صعود التدرج العشوائي. في حين تقوم أساليب تدرج السياسات التقليدية بإجراء تحديث تدرج واحد لكل عينة بيانات، فإن PPO تقترح دالة هدف جديدة تمكن من إجراء عدة حقب من تحديثات الدفعات الصغيرة. وذلك من خلال محاولة حساب تحديث في كل خطوة تقلل من دالة التكلفة مع ضمان أن يكون الانحراف عن الدالة السابقة صغير نسبيا. لذا تقوم بأمثلة السياسة المدربة بشكل حذر، عن طريق تقليل الفروقات المفرطة بين السياسة القديمة والجديدة باستخدام دالة هدف مقطوعة، مما يمنع تدهور الأداء أو حدوث تعلم غير مستقر. غالبا تنفذ الخوارزمية بنهج الفاعل – الناقد الذي يعتمد على التشارك بين عنصر الفاعل الذي يحدد الأفعال، وعنصر الناقد الذي يقيم جدواها. ويتم تحديث كليهما بشكل متزامن في كل تكرار تدريبي. تتضمن العملية تخزين التجارب التفاعلية بين الوكيل والبيئة. ثم حساب تقدير المنفعة Advantage Estimate لكل خطوة زمنية، يليها أمثلة تدريجية للسياسة. ويتم حساب دالة الهدف المقطوع L^{clip} ثم أمثلة السياسة بطريقة حذرة. يتم في الأسطر التالية عرض خوارزمية PPO بأسلوب الفاعل الناقد حيث يتم تنفيذ الخوارزمية باستخدام التكرارات التدريبية التالية:

for iteration = 1, 2, 3, … **do**
  for actor = 1, 2, 3, …, N **do**
    Run Policy $\pi_{\theta\ old}$ in environment for T time steps
    Compute advantage estimates $\hat{A}_1, \ldots . \hat{A}_T$
  end for
  Optimize surrogate L wrt Θ, with K epochs and minibatch size $M \leq NT$
  $\theta_{old} \leftarrow \theta$
end for

حيث، في كل تكرار يتم تشغيل السياسة القديمة لمدة T خطوة زمنية، ويجمع N فاعل على التوازي البيانات. ثم يحسب تقدير الميزة، ويليه حساب دالة الهدف البديل L على هذه الخطوات NT وأمثلتها باستخدام minibatch size SGD أو أحيانا محسن آدم، لمدة K حلقة. حيث لكل تكرار (دورة تدريبية كاملة) يطلب من الوكيل (نسخة من المتعلم) تشغيل السياسة الحالية في البيئة لمدة T خطوة زمنية وتخزين بيانات التفاعل (Action, State, Rewards)، ثم حساب المنفعة المقدرة التي تقيس جودة الإجراء مقارنة بمتوسط متوقع لكل خطوة بعد انتهاء جمع التجارب لـActor وحساب أفضلية الإجراء يتم أمثلة السياسة القديمة بطريقة حذرة، ويتم تكرار هذه الدورة عدة مرات حتى يتحسن أداؤه تدريجيا (Schulman et al, 2017) (Ugurlu, 2022).

4.3.10.التعلم العميق DL – Deep Learning:

غيّر التعلم العميق الطريقة التي نعالج بها البيانات ونحللها ونتعامل معها وقدم طرقا جديدة للتعامل مع البيانات ومعالجتها وتحليلها. ويحقق أحيانا نتائج مماثلة لأداء الخبراء البشريين أو متفوقة عليه، وأصبح شائعا جدا في عائلة تعلم الآلة مع أدائه المتميز في مجموعة متنوعة من المجالات ليس فقط في مهام الرؤية الحاسوبية الكلاسيكية ولكن أيضا في العديد من التطبيقات العملية الأخرى؛ على سبيل المثال، معالجة اللغة الطبيعية، والترجمة الآلية، وعلم المعلومات الحيوية، وألعاب الطاولة، وغيرها (Liu et al, 2021). وتأتي صفة "عميق" من استخدام طبقات متعددة في الشبكة. يتم إجراء مقارنة عامة في الجدول (2–4) بين التعلم العميق وتعلم الآلة التقليدي. مع التعلم العميق، يتم تغذية البيانات الخام مثل الصور مباشرة إلى طبقات متعددة من الشبكة العصبية العميقة التي تستخرج تدريجيا ميزات أعلى مستوى، بينما مع تعلم الآلة التقليدي، يتم استخراج الميزات عالية المستوى ذات الصلة لبيانات الإدخال يدويا بواسطة الخبراء. علاوة على ذلك، يتطلب التعلم العميق في كثير من الأحيان كمية كبيرة من البيانات للوصول إلى نتائج مثالية، وبالتالي فهو يتطلب أيضا قدرا كبيرا من العمليات الحسابية. ويسند أول نجاح كبير للتعلم العميق إلى شبكة عصبية ملتوية للتصنيف في عام 2012. على الرغم من أن هذه التقنية قد تم تحسينها باستمرار وبسرعة منذ انطلاقها وهي الآن واحدة من أكثر هياكل تعلم الآلة شيوعا إلا أنها ليست مناسبة تماما للتحكم في التلاعب الروبوتي حيث يستغرق الأمر وقتا طويلا للحصول على عدد كبير من صور زوايا المفاصل أو المدخلات الكثيرة مع بيانات مُسمّاة لتدريب النموذج. في الواقع، هناك بعض الأبحاث التي تستخدم الشبكة العصبية التلافيفية لمعرفة عزم دوران المحرك اللازم للتحكم في الروبوت باستخدام صور فيديو ملونة خام. ومع ذلك، فإن الفكرة الأكثر وعد وإثارة للاهتمام هي استخدام DRL كما سنناقش في الفقرة التالية (Liu et al, 2021).

الجدول 2–4 مقارنة التعلم العميق بالتعلم التقليدي (Liu et al, 2021):

	تعلم الآلة التقليدي	التعلم العميق
متطلبات مجموعة البيانات	أداء جيد مع مجموعات صغيرة	تحتاج مجموعة بيانات ضخمة
الدقة	ثبات الدقة	احتمال أداء ممتازة
استخراج السمات	اختيار يدوي	تعلم تلقائي
بنية الخوارزمية	نموذج بسيط	نموذج متعدد الطبقات
وقت تدريب النموذج	سريع	كثيف حسابيا

4.3.11.التعلم المعزز العميق Deep Reinforcement Learning – DRL:

التعلم المعزز العميق هو دمج RL مع DL، فهو يجمع بين تقنية منح المكافآت بناءً على الإجراءات من RL وفكرة استخدام شبكة عصبية عميقة لتعلم تمثيلات الميزات من DL. يقتصر RL التقليدي على المجالات ذات التمثيلات البسيطة للحالة بينما يجعل DRL من الممكن للوكلاء اتخاذ القرارات من بيانات الإدخال عالية الأبعاد وغير المنظمة باستخدام الشبكات العصبية لتمثيل السياسات. في السنوات القليلة الماضية، كان البحث في DRL نشطا للغاية مع قدر كبير من التقدم، جنبا إلى جنب مع الاهتمام المتزايد بالتعلم العميق. بالنسبة للمشكلات البسيطة من الممكن أن نخزن جميع قيم الحالة أو قيم إجراءات الحالة في جدول، ونحسبها واحدة تلو الأخرى. بالنسبة للمشكلات الأكثر تعقيدا التي يجب أن نواجه فيها لعنة الأبعاد، يمكن أن يكون عدد الحالات كبيرا جدا بحيث لا يمكن معالجتها باستخدام أنظمة الكمبيوتر الأساسية. مثل حالة يد إنسان آلي بخمس أصابع، لكل منها اثنين إلى ثلاثة مفاصل، يتمثل أحد الحلول للتغلب على هذه المشكلة في استخدام طرق تقريب الوظيفة الخطية أو غير الخطية أحيانا مثل الإجراءات الخطية أو أشجار القرار أو الجيران الأقرب أو الشبكات العصبية العميقة. الشبكات العصبية الاصطناعية قادرة على التعميم من بيانات التدريب للتعامل مع البيانات غير المرئية. يمكن استخدام الشبكة العصبية للتنبؤ بقيم إجراءات الحالة لأزواج إجراءات الحالة غير المرئية، إذا تم تدريبها أولا على العديد من الأمثلة. يسمح هذا الجمع بحل القضايا الحرجة المتعلقة بأبعاد البيانات وقابليتها للتوسع في المهام ذات إشارات المكافأة مثل التلاعب الروبوتي ومهام التحكم، والتي لا تسمح أي من الطريقتين بحلها عند تطبيقها بمفردها (Liu et al, 2021) (Kepes, 2017).

4.4. تعلم المناهج Curriculum Learning - CL:

أسلوب آخر لتعلم الآلة مستوحى من علم الأحياء، حيث أن الأحياء يمكنهم أن يتعلموا بشكل أكثر كفاءة إذا تم تنظيم مادة التدريب بطريقة هادفة وبنية جيدة وأُدخل التعقيد تدريجيا. تم استخدام هذا في تعلم الآلة من خلال تعلم المناهج كاستراتيجية أمثلة شاملة، هذه التقنية تجد القيم الدنيا المحلية بشكل أفضل بل وتتقارب مع الحد الأدنى العالمي بشكل أسرع من الطرق الأخرى وفي هذه التقنية تزداد صعوبة الأهداف تدريجيا مع تعلم الوكيل (kepes, 2017).

4.5. أهمية المحاكاة في تطبيقات التعلم المعزز:

أصبحت المحاكاة في السنوات الأخيرة جزءا أكثر أهمية من أي وقت مضى في تطوير الأجهزة وبشكل خاص مجال الروبوتات وRL. فكما رأينا في فصل الدراسات السابقة تعتمد بكثرة على المحاكاة إذ يتم البدأ بتدريب الوكلاء ضمن المحاكاة ثم يُنتَقل للتجارب الفيزيائية، وتعتمد هذه المنهجية في دراسات الروبوتات وRL لأن الأخير يبدأ بتجارب تكون في البداية عشوائية جدا وخاطئة، إذا بدأ بها على الروبوتات الحقيقية بشكل فوري ستقود لتلف الروبوت ويصبح الموضوع مكلف ومرهق جدا وغير عملي. تحتاج خوارزميات RL على غرار أساليب تعلم الآلة الأخرى

لكمية بيانات كبيرة للتدريب على المهمة يقوم بجمعها آنيا من خلال مبدأه المعتمد على التجربة والخطأ، وهي أحد أصعب المهام وأكثرها تكلفة. ومن هنا ظهرت الحاجة لبرامج المحاكاة الفيزيائية التي تسمح للوكلاء بجمع الكثير من الخبرة من خلال المحاكاة وعلى عكس عمليات المحاكاة المستخدمة في طريقة العناصر المحدودة (Finite Element Analysis - FEM) وديناميكيات السوائل المحوسبة (Computational Fluid Dynamics - CFD)، والتي يمكن أن تستغرق ساعات أو أياما لخطوة محاكاة واحدة، تتطلب التطبيقات الروبوتية وRL بيئة محاكاة سريعة الاستجابة تعمل على الأقل في الوقت الفعلي. يمكن تحقيق هذه الاستجابة عندما يكون عامل الوقت الحقيقي وهو (ratio of simulation time over real-world time - RTF) نسبة وقت المحاكاة عبر الزمن الحقيقي على الأقل حوالي 1. تستفيد تطبيقات DRL بشكل خاص عندما يكون RTF>1 حيث يمكن تدريب الوكيل بشكل أسرع من الوقت الفعلي مما يؤدي إلى تقليل أوقات الانتظار للتدريب. في حين أن المحاكاة الروبوتية الرقمية البحتة بشكل عام لا تستفيد من RTF أعلى من 1، فإن DRL تستفيد بكثرة. إن هذه الزيادة غالبا ما تؤثر على الدقة فتعطي دقة أقل وبالتالي نظرا لأن FEM وCFD تحتاج إلى أن تكون دقيقة قدر الإمكان فهي بطيئة وفقا لذلك. لا يحتاج التطبيق الآلي وDRL في معظم الحالات إلى مثل هذه الدقة، وبالتالي يمكنه استخدام بيئات المحاكاة ذات الدقة المنخفضة لصالح السرعة. تتمثل إحدى الطرق البسيطة والمستخدمة غالبا لزيادة السرعة في زيادة الخطوة الزمنية للمحاكاة مع تقليل الدقة في نفس الوقت. ومع ذلك يجب أن تكون الدقة جيدة بما فيه الكفاية مما يعني أن المحاكاة يجب أن تصور فيزياء العالم الحقيقي بشكل جيد بما يكفي لحالة الاستخدام الفردي للعمل. بشكل عام، يجب النظر في التوازن بين الدقة والاستقرار والأداء، ورغم أن المحاكاة تهدف إلى محاكاة الحقيقة، إلا أن لها العديد من المزايا والعيوب مقارنة بالتجارب الفيزيائية (Sherstan, 2015) (Sun et al, 2021) (Korber et al, 2021) (Kepes, 2017).

<u>إيجابيات المحاكاة:</u>

1) يمكن تشغيل المحاكاة بسرعة أكبر.
2) يمكن ضمن المحاكاة التعامل مع الأخطاء والحوادث بشكل مريح فلا تؤدي لنفقات كبيرة، مثلا يعتبر تعليم الروبوت على الإمساك مع DRL أكثر خطورة في الحياة الواقعية حيث يمكن أن ينكسر الروبوت أو الجسم الملتقط بسبب الفشل المتكرر في بداية استكشاف البيئة لذا يفضل دائما البدء مع المحاكاة والتطوير ثم النشر.
3) تسهل المحاكاة إعادة ضبط البيئة باستخدام عملياتها، على سبيل المثال يمكن إعادة ضبط الجسم الملتقط بعد انتهاء التجربة إلى مكانه الأصلي تلقائيا في حين يكون هذه الجزء مستهلك للوقت كثيرا في التطبيقات الواقعية.
4) يمكن تشغيل عمليات محاكاة متعددة معا مما يؤدي لتعلم أسرع وهو بالطبع ممكن في الواقع لكن بتكلفة كبيرة.
5) من الممكن الوصول للمحاكاة في أي وقت من اليوم وهو ما لا يحدث عادة مع الروبوت الحقيقي.

سلبيات المحاكاة:

1) حتى أفضل عمليات المحاكاة يتم تبسيطها ولا يمكنها أخذ جميع القيود الواقعية في الاعتبار. وبالتالي لتسريع عمليات المحاكاة يتعين علينا غالبا التضحية بقدر أكبر من الدقة.
2) المتابعة والنشر، رغم أن الكثير من الدراسات بينت أنه يمكن نشر نماذج RL بنجاح من المحاكاة إلى الواقع إلا أنه ليس من المضمون أن يعمل النموذج في الواقع.

4.6.مفهوم فجوة الواقع:

يلعب نقل التعلم من المحاكاة إلى الواقع دورا حاسما في التطبيقات الروبوتية. وقد قدمت العديد من الدراسات الحديثة نتائج واعدة في هذا المجال. السؤال الأهم في قضية المحاكاة هو: بمجرد تدريب النموذج على المحاكاة، هل من الممكن نشره بسهولة على الأنظمة المادية؟ في الواقع، إذا لم يتم اتخاذ احتياطات معينة، فإنها لا تعمل. وهذا ما يسمى فجوة المحاكاة إلى الواقع. الفجوة بين المحاكاة والنظام الحقيقي من حيث الديناميكيات والإدراك تسبب انخفاضا قد يكون كارثيا في الأداء معظم الأحيان، حتى المهمة البسيطة للغاية التي يتم تعلمها جيدا في المحاكاة لن يتم تنفيذها بشكل صحيح على الأنظمة الفيزيائية. ستسمح الفجوة الضيقة بين المحاكاة والحياة الواقعية بتدريب روبوت محاكى ومن ثم نشر نموذج RL المدرب في منصة تحكم آلية حقيقية. ولمعالجة هذه المشكلة، غالبا ما يتم استخدام طريقتين. الأولى زيادة جودة المحاكاة للاقتراب قدر الإمكان من الواقع. تتمثل الطريقة الثانية في جعل وحدة التحكم قوية في مواجهة التغييرات في خصائص النظام. إذا كانت وحدة التحكم قوية لتغيرات النظام فستكون أيضا قوية لتغيرات النظام عند الانتقال من المحاكاة إلى الواقع، مما يسمح بنشر أفضل. يمكن استخدام التوزيع العشوائي لبعض المتغيرات في المحاكاة لتحقيق هذه المتانة. مثل، ترتيب ديناميكيات النظام بشكل عشوائي، أو استخدام سياسة عشوائية أو إضافة ضوضاء إلى الملاحظات. عمليا، تستخدم الطريقتين معا (Gallouedec et al, 2020).

4.7.الخاتمة:

شكل التعلم المعزز العميق نهجا واعدا في مجالات التحكم بالأذرع والأيدي الروبوتية ولكن لا يزال حتى اليوم هناك حاجة كبيرة لتكثيف الدراسات حول هذا الموضوع، وبهدف تطبيق هذا النهج ضمن البحث الحالي تمت مراجعة تاريخ هذا المجال وأهم مفاهيمه التي لزمت لإنجاز هذا البحث.

5. الفصل الخامس: منهجية البحث.

5.1. تمهيد:

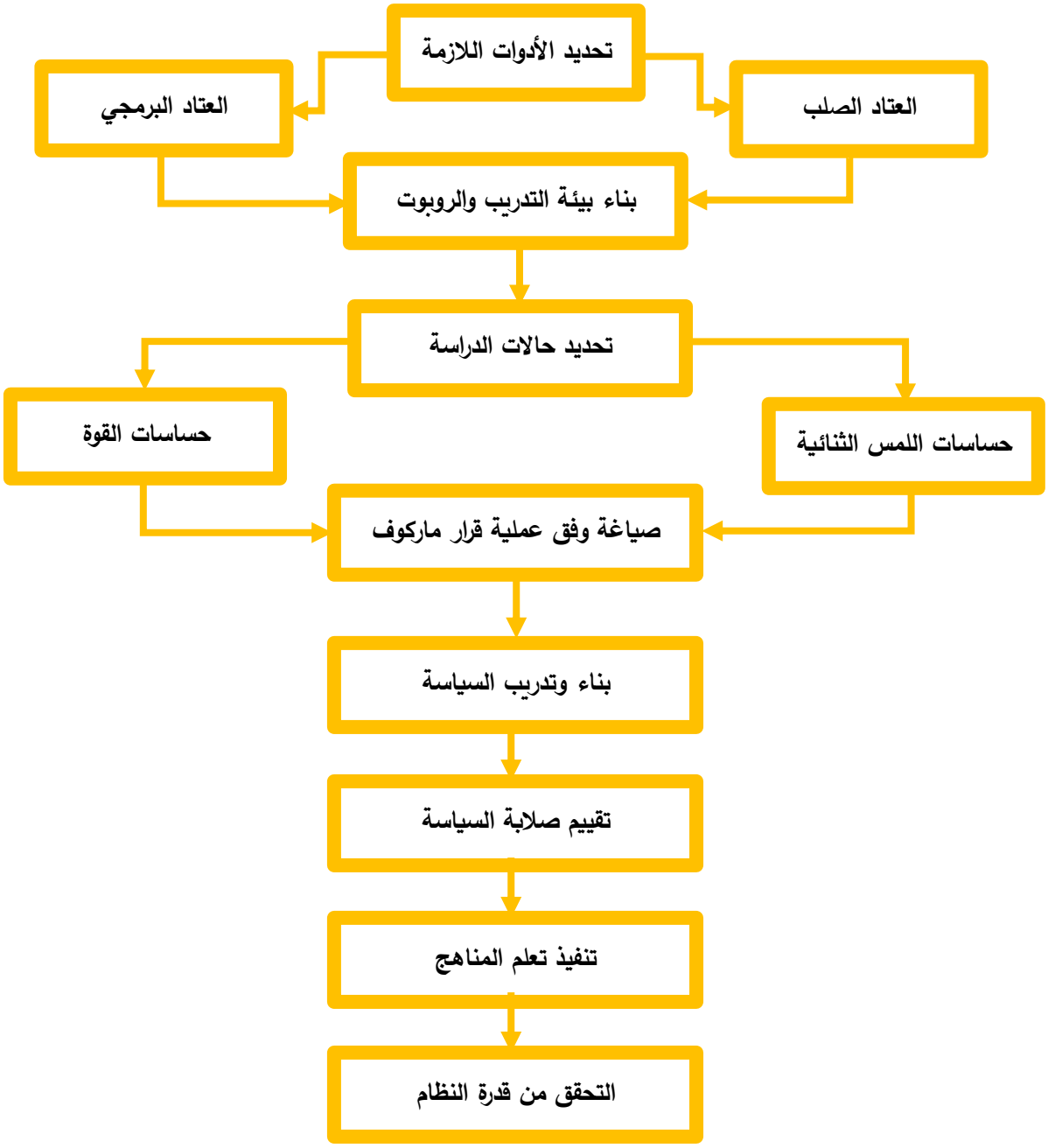
يركز هذا البحث على تدريب وكيل ذكي متمثل بيد طرف اصطناعي بخمسة أصابع لاكتساب سياسة تحكم فعالة تجعل اليد الروبوتية قادرة على تنفيذ مجموعة من القبضات المتنوعة. وقد تم اتباع منهجية منظمة لتحقيق النتائج المرغوبة، يعرض في هذا الفصل منهجية البحث ومراحله بشكل منسق ومتتابع خطوة بخطوة.

5.2. منهجية البحث:

تتكون المنهجية المتبعة لتنفيذ البحث من عدة خطوات، يعرض الشكل (1-5) منهجية العمل بشكل مرئي متكامل، ويوضح تسلسل الخطوات الرئيسية منذ صياغة المشكلة وحتى اختبار السياسة النهائية، وهي:

1) تحديد المواد والأدوات اللازمة لتنفيذ الجزء العملي من البحث.
2) بناء بيئة التدريب، التي سيتدرب ويختبر لاحقا ضمنها الوكيل.
3) تحديد حالات للدراسة، وقد تم تحديد حالتين لدراستهما وهما: (الأولى هي مع حساسات اللمس الثنائية. والثانية هي مع حساسات القوة) ولكل حالة درست الخطوات (4 حتى 8) التالية:
4) صياغة المشكلة وفق نموذج عملية قرار ماركوف، حيث تم تمثيل بيئة التعلم كوصف رياضي يعتمد على نموذج عملية قرار ماركوف، بحيث تكون كل حالة مرتبطة بإجراء معين ومكافأة، مما يتيح بناء نموذج تفاعلي قابل للتدريب باستخدام تقنيات التعلم المعزز العميق.
5) بناء وتدريب السياسة: تم في هذه المرحلة تنفيذ تجارب مكثفة لتحديد السياسة المثلى، شملت اختبار عدة تركيبات من محددات التدريب مثل (معامل القطع، ومعدل التعلم، وطول الحلقات، وحجم العينة، وبنية الشبكات في السياسة، وما إلى ذلك) إلى جانب تقييم توابع مكافأة ومساحات إجراءات مختلفة. بمجرد تحقيق أداء مرض يتجلى بزيادة قيمة المكافأة التراكمية إلى قيمة مرغوبة يتم تعيينها عن طريق التجربة والخطأ، تم تحديد تابع المكافأة الأنسب، والمنهجية العامة لبناء سياسات الإمساك، ثم تم اعتماد السياسة الأفضل وإيقاف عملية التدريب، والانتقال لمرحلة الاختبارات.
6) تقييم صلابة (Robustness) السياسة المدربة: تمت دراسة مدى قدرة سياسة اتخاذ القرار (Policy) النهائية الناتجة عن المرحلة السابقة على التعامل مع مواقف جديدة أو بيئات غير مألوفة، ويقصد بصلابة السياسة قدرة الوكيل على الاستمرار بالنجاح رغم تغيّر الشروط أو ظهور عقبات لم يدرب عليها بشكل مباشر، وذلك من خلال تجارب اختبارية مصممة بعناية لهذا الغرض.
7) مرحلة تعلم المناهج للاستفادة من التعلم التدريجي، تم اعتماد منهجية "تعلم المناهج" بحيث يتم تدريب الوكيل على مهام بسيطة بداية، ثم تعقيدها تدريجيا بما يتيح للوكيل بناء قدرات متقدمة بشكل تراكمي، مما يسهل الوصول إلى سياسة أكثر تعميما وكفاءة.
8) التحقق من قدرة النظام المقدم (البنية المقترحة وهيكل المكافأة النهائي ومساحات الإجراءات) على إنشاء قبضات مخصصة مستقبلا: تم اختبار فعالية المنهجية المقترحة من خلال إعادة تدريب وكيل جديد من نقطة الصفر على مهام قبض جديدة. تهدف هذه الخطوة إلى تقييم قدرة هذا الهيكل والنهج

المقدم على توجيه سلوك الوكيل نحو نتائج مرضية في بيئات جديدة لم يتعرض لها مسبقا، بحيث تُقييم قدرة هذا النهج على التدريب مستقبلا لإنشاء قبضات خاصة وملائمة لنوعية حياة المبتور.



الشكل 1-5 منهجية العمل.

5.3. تحديد أدوات البحث:

يعتمد البحث على المحاكاة التي تحتاج قدرات حاسوبية مرتفعة وبرامج ذات وثوقية عالية، ولتنفيذه عمليا وتحقيق النتائج المرجوة بحيث تكون دقيقة وموثوقة، كان من الضروري اختيار كل من العتاد الصلب والبرمجيات بعناية فائقة، أخذت هذه الخطوة جهد ووقت كبيرين، لضمان دقة المحاكاة وفعالية التحليل الحاسوبي للبيانات لاحقا.

5.3.1. العتاد الصلب:

5.3.1.1. الحاسوب المستخدم:

تُجرى كامل تجارب هذا البحث في بيئة المحاكاة، حيث يمكن تدريب النماذج واختبارها في ظل مجموعة واسعة من الخصائص الفيزيائية كشروط الحمل والاحتكاك والكثافة وغيرها، ويتطلب هذا أجهزة حاسوبية قوية قادرة على تحمل عبء العمل الحسابي كما تتطلب المحاكاة أيضا موارد رسومية عالية، وبالتالي وحدة معالجة مركزية Central Processing Unit - CPU وكذلك وحدة معالجة الرسوميات Graphics Processing Unit - GPU قوية أيضا. في هذا البحث تم اختيار حاسوب HP PROBOOK له المواصفات المعروضة في الجدول (1-5). ومن الجدير بالذكر أن جميع تجارب واختبارات المحاكاة أجريت أثناء التوصيل بالتيار الكهربائي ويتم تبريد المعالج أثناء وبعد كل استخدام وتجربة لأن الدراسات التي تمت مراجعتها أوصت بهذه الخطوة لزيادة موثوقية النتائج، والتي تبين بالتجريب العملي تأثرها عند فصل التيار الرئيسي أو ارتفاع حرارة المعالج.

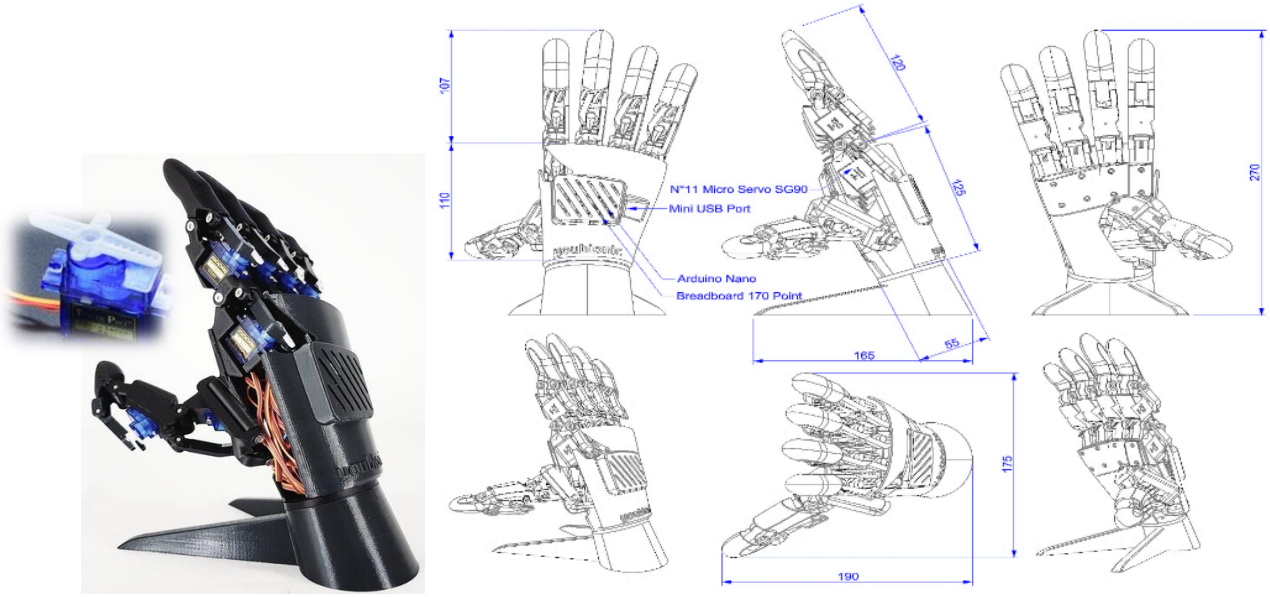
الجدول 1-5 خصائص الحاسوب الذي ستجرى عليه الدراسة.

Name	CPU	RAM	GPU	STORAGE	OPERATING SYSTEM
HP PROBOOK 455 G7	AMD Ryzen 5 4500U 2.38 GHz	8 GB	AMD Radeon Graphics	SSD 250 GB	WINDOWS 10 (64 BIT)

5.3.1.2. اليد التي أجريت عليها التجارب:

أُجريت أبحاث كثيرة حول التصميم البنيوي للأيدي الاصطناعية التي يمكن من خلالها تقديم حركات تحكم دقيقة. ومع تحسن الإلكترونيات والموارد الحاسوبية، تتطور تصميمات أنظمة اليد للأطراف الاصطناعية للتحكم الآلي الدقيق تدريجيا من درجات حرية أقل إلى درجات حرية أكثر، ومن الصلابة إلى المرونة، ومن عدم الإحساس إلى الاندماج المتعدد الاستشعار. قامت الدراسات السابقة بالتدريب على نماذج عامة للأيدي الروبوتية، في هذا البحث وسعيا لتقليص الفجوة بين المحاكاة والواقع تم تصميم نموذج اليد ضمن المحاكاة ليكون أقرب ما يمكن من حيث الصفات الميكانيكية والفيزيائية للواقع بحيث يمكن متابعة التجارب في البيئة الواقعية لاحقا بسهولة عليه، لذا أخذت هذه الخطوة وقت طويل نسبيا لاختيار طرف حقيقي ذو مواصفات ودقة مميزة وسهل التصنيع ويمكن بسهولة استخدامه. في النهاية تم اختيار نموذج يد HANDY الذي وفرته شركة YOUBIONIC الإيطالية عام 2019، التي تطور الروبوتات وأسسها المهندس في علم الروبوتات Federico Ciccarese، تتيح الشركة ملفات الطرف المختار بالطلب عبر الإيميل youbionic@gmail.com بسعر 200 دولار (لليد) إلى 300 (لليد مع ذراع). وتشمل الملفات (ملفات STL يمكن طباعتها مباشرة أو تعديلها، وينصح بمادة PLA التي تكسب الطرف المقدم قوة ومرونة في ذات الوقت، وملفات الدارة الكهربائية للتوصيل وكود برمجة، وملفات الرسم الفني، ودليل التركيب). عادة لا

يتجاوز وزن اليد نصف كيلو غرام مع تعبئة 100%. وتتميز بتقنيات متطورة وقدرات حركة غير محدودة، وتكون قادرة على أداء جميع أنواع القبضات والوضعيات بفضل محرك مؤازر مثبت على كل مفصل فيها والذي يمنحها 11 درجة من الحرية تمنحها القدرة على تشكيل كم واسع من القبضات، إذ تملك اليد محركين مؤازرين لكل إصبع، ومحرك مؤازر إضافي في الإبهام يتولى تدويره نحو باطن الكف، ليصبح المجموع 11 محرك مؤازر نوع MICRO SERVO SG90. وبحسب الشركة تم استخدام اليد من قبل العديد من المبتورين في الأطراف العلوية، وعلى عكس الاستخدامات المحدودة للأطراف المتوفرة الآن، فإنه يمكن أن تشمل مساحات استخدام هذه اليد، (أداء المهام الروتينية، تجربة تطبيقات الذكاء الاصطناعي، التحكم عن بعد). يوضح الشكل (2-5)، نموذج اليد المختار وأبعاده ويوضح مقاربته لليد البشرية من حيث الحجم وعدد الأصابع. وتتميز هذه الذراع بدقة عالية وتعدد الاستخدامات ويمكن رؤية بعض من نطاق الحركة الذي تقدمه، فيديو للذراع كاملة: https://youtu.be/71dJv20Ir3U?si=x5DE7tx0r_IUBwvK. فيديو لاستخدام اليد كطرف اصطناعي: https://youtu.be/hLZUebOZwSQ?si=t6tOf5iSi2z3Jy0. فيديوهات لليد: https://www.youtube.com/watch?v=XM6gM7bLmcs .https://youtube.com/shorts/Yr4KHEZuTFo?si=9dR9_D3DOfEqgcTc



الشكل 2-5 نموذج وأبعاد يدHANDY

5.3.2. العتاد البرمجي:

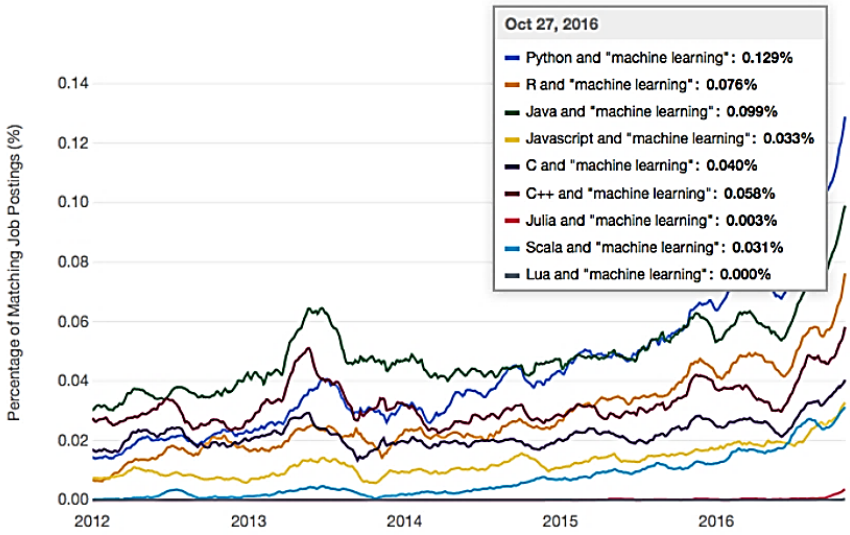
5.3.2.1. برنامج المحاكاة:

ترتبط جودة بيانات التدريب ارتباط وثيق بجودة بيئة المحاكاة التي تولدها. كما تعتمد كمية بيانات التدريب على سرعة بيئة المحاكاة. ومع تزايد تطبيقات DRL ظهرت عدة برامج محاكاة وعدة أبحاث تركز على قدرات المحركات الفيزيائية (Open Dynamics Engine - ODE) وتقوم بمقارنتها لتسهل على الباحثين اختيار بيئة موثوقة مع نموذج فيزيائي يمثل البيئات بشكل جيد بما يكفي لإنتاج بيانات ذات مغذى وبسرعة مقبولة. قام Korber وزملاؤه عام 2021 بمقارنة أربع من أشهر بيئات المحاكاة الشائعة في مجتمع تطبيقات الروبوتات والتعلم المعزز العميق

وهي، Gazebo وMuJoCo وPyBullet وWebots. من خلال سلسلة معايير لتسهل اختيار بيئات المحاكاة على باحثين DRL والذين يحتاجون إلى بيئات محاكاة سريعة ومستقرة ودقيقة، عن طريق تقديم تقييمات لاستقرار المحاكاة مقابل سرعتها في التنفيذ، والاستهلاك الحسابي. بينت نتائج الدراسة أنه إذا كانت المهمة تتطلب الدقة والثبات، فإن Webots هو الخيار الأفضل (حيث أبدى استقرارا جيد في الاختبارات، حتى خطوة زمنية تصل إلى 48 مللي ثانية، مع الاحتفاظ بدرجة عالية من RTF)، خاصة في المهام التي تتطلب التحكم الدقيق مثل التكديس أو التحكم الروبوتي. أما إذا كانت المهمة تتطلب محاكاة سريعة للتصادمات المعقدة، فقد يكون Gazebo أو MuJoCo أفضل (Korber et al, 2021). وبناء على نتائج هذه الدراسة وتوصياتها ونظرا لاستخدامه في العديد من الدراسات الحديثة مثل: (Trararak et al, 2025) (Tao et all, 2023) (Coskun et al, 2022) (Gomes et al, 2022). تم استخدام برنامج Webots، الذي طورته شركة Cyberbotics في هذا البحث. والذي يتميز بدعمه الكامل لمحاكاة الروبوتات وخوارزميات RL، ويعد أكثر حداثة وواقعية من البيئات الأخرى الشائعة في تطبيقات DRL (Kirtas et al, 2020). وهو حزمة برمجية احترافية لمحاكاة الروبوتات ويوفر بيئة نماذج أولية سريعة لنمذجة وبرمجة ومحاكاة الروبوتات. تم تطويره واختباره وتوثيقه بدقة لأكثر من 25 عام. يقدم برنامج Webots مكتبة واسعة من نماذج العوالم وأجهزة الاستشعار والروبوت. يمكن برمجة الروبوتات فيه بلغة C أو C++ أو Python أو Java أو MATLAB أو ROS مع وثائق API مفصلة لكل خيار . يمكن استيراد النماذج من تنسيقات CAD متعددة أو URDF. يستخدم في جوهره محرك ODE مخصصا كمحاكي افتراضي للفيزياء يعطيه دقة في الفيزياء ومعالجة الاحتكاك والتفاعل بين الأجسام (Korber et al, 2021). ويوفر بيئات قابلة للتخصيص والقدرة على إنشاء روبوتات من البداية بالإضافة لعمليات محاكاة عالية الدقة برسومات واقعية (Kirtas et al, 2020). لكل كائن يمكن تحديد عدد من الخصائص كالشكل واللون والملمس والكتلة والاحتكاك وما إلى ذلك. يمكن تزويد كل روبوت بعدد كبير من أجهزة الاستشعار والمحركات المتاحة. كما يتيح نقل وحدات التحكم إلى روبوتات حقيقية ومحاكاة أنظمة متعددة الوكلاء. يمكن أن تعمل المحاكاة بسرعة تصل إلى 300 مرة أسرع من الحقيقية عند الوضع السريع وتُضبط خطوة وقت المحاكاة لتناسب الاحتياج (الدقة مقابل السرعة) كما يوفر وضع خطوة بخطوة لدراسة تصرف الروبوت بالتفصيل. يرتبط كل مكون من مكونات الروبوت بجسم للكشف عن التصادم أثناء المحاكاة. يمكن التحكم في المحركات بواسطة برامج تحكم في عزم الدوران أو الزاوية أو السرعة. تحتوي شجرة المشهد المهيكلة كملف VRML97 على جميع المعلومات اللازمة لوصف المحاكاة وتتألف من قائمة من العقد كل منها تحتوي على حقول ويمكن أن تحتوي الحقول على قيم أو عقد. يعمل على أنظمة التشغيل Windows وLinux وMac OS X. مخصص للباحثين المهتمين بالروبوتات وتم استخدامه من قبل مئات الجامعات والمركز البحثية في جميع أنحاء العالم منذ عام 1998 حتى الآن لأغراض التعليم والبحث (Michel, 2004).

5.3.2.2. لغة البرمجة:

يوجد عدة لغات برمجة يمكن لبرنامج Webots التعامل معها لتدريب وكيل DRL، في هذا البحث سيتم التعامل مع لغة برمجة PYTHON 3.7 إذ تعد من أكثر لغات البرمجة شيوعا في مجال ML كما يوضح الشكل (3-5)، وتدعم العديد من الأطر التعليمة والمكاتب مثل NumPy، Keras، وما إلى ذلك (Kepes, 2017). وتم استخدام هذه النسخة من بايثون لأنها متوافقة مع إطار DEEPBOTS الذي سنتحدث عنه في الفقرة التالية.

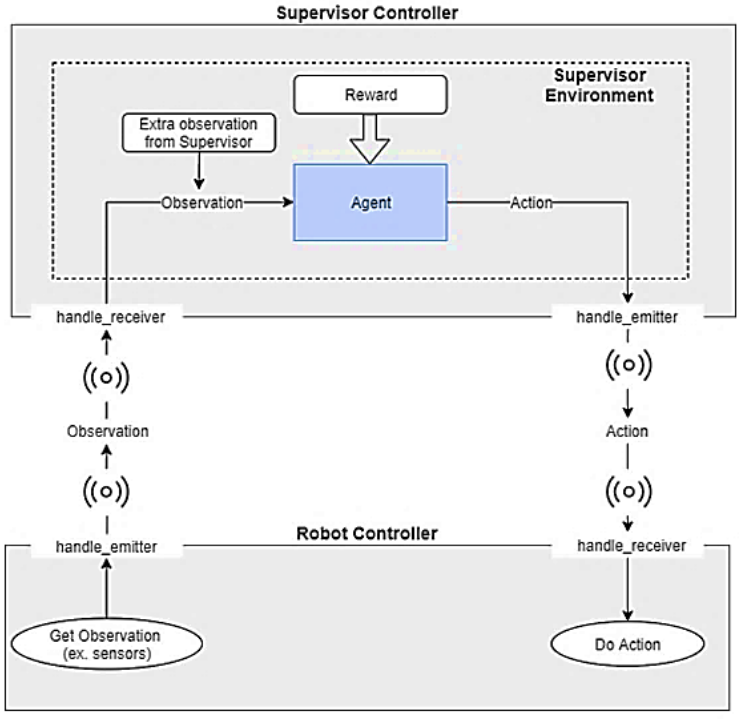


الشكل 3-5 أكثر لغات البرمجة انتشار في مجال التعلم المعزز (Kepes, 2017).

5.3.2.3. إطار DEEPBOTS:

رغم القوة التي وفرتها المحاكيات الحديثة للروبوتات إلا أنه لا يكون عادة من السهل استخدامها لتطوير مسائل DRL التي تتطلب مستوى أعلى من التجريد يخفي التفاصيل منخفضة المستوى مثل كيفية التحكم الفعلي ومعالجة الأوامر بواسطة الروبوت. بالإضافة لمتطلبات التعلم الكبيرة والمقدار الهائل من خبرات التطوير التي تحتاجها. مما حدَّ من فائدة استخدام المحاكيات في تطوير مسائل DRL من قبل الباحثين. للتغلب على هذه القيود قدم Kirtas وزملاؤه إطار عمل مفتوح المصدر سمي DeepBots يجمع بين واجهة سهلة لمطوري DRL، وواجهة OpenAI Gym المعروفة، مع أحدث محاكي للروبوتات وهو Webots، من أجل توفير طريقة قياسية لاستخدام DRL مع الروبوتات. تهدف Deepbots إلى تمكين الباحثين من تطوير أساليب DRL بسهولة ضمن Webots من خلال تقليل جهود التطوير المطلوبة. ويأتي بشكل حزمة بايثون سهلة التثبيت تسمح بتنفيذ بيئات متنوعة بكفاءة لاستخدامها من قبل الباحثين، والمطورين لمجموعة الخوارزميات الخاصة بهم. بالإضافة إلى ذلك تتيح الحزمة أدوات مراقبة للتدريب. يوضّح الهيكل الأساسي لنظام DeepBots في الشكل (4-5)، حيث تكون وحدة تحكم المشرف وهي الكود الخاص بالمشرف ووحدة التحكم في الروبوت وهي كود الروبوت. فمن المعروف عند تنفيذ RL في البيئات المختلفة من الضروري تضمين عقدة مشرف خاصة لديها معرفة كاملة بالبيئة ويمكنها الحصول على معلومات

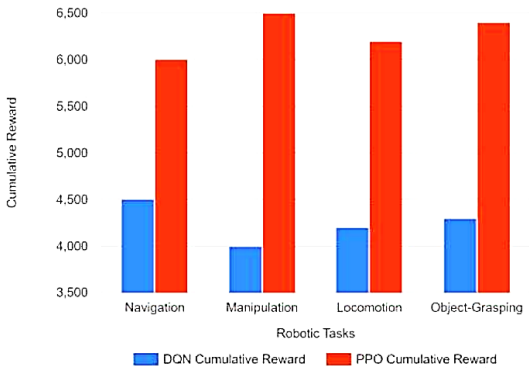
وتعديل كل العقد الأخرى. يتم الاتصال بين المشرف والروبوت عبر بواعث/مستقبلات، تبث وتستقبل الرسائل على التوالي. عادة يكون المشرف عقدة بدون كتلة أو أي خصائص فيزيائية في المحاكاة. على سبيل المثال، يمكن أن يكون المشرف جهاز كمبيوتر محمول ينقل الإجراءات إلى الروبوت، ولكن دون التفاعل مع المشهد الفعلي. علاوة على ذلك، يمكن أن يكون المرسل والمستقبل أي جهاز محتمل، سواء كان سلكي أو لاسلكيا. تعمل Deepbots على النحو التالي: أولا وقبل كل شيء، يجب إعادة ضبط جهاز المحاكاة في الحالة الأولية. من ناحية، يقوم الروبوت بجمع المجموعة الأولى من الملاحظات وباستخدام الباعث الخاص به يرسل المعلومات إلى عقدة المشرف. من ناحية أخرى، يتلقى المشرف المشاهدات من مكون المستقبل الخاص به ويقوم بدوره بتمريرها إلى الوكيل. في هذه الخطوة، إذا لزم الأمر، يمكن للمشرف زيادة الملاحظة بمعلومات إضافية، كالمسافات الإقليدية مثلا. باستثناء المشاهدات، يمكن للمشرف تمرير الإجراء السابق والمكافأة و/أو أي معلومات حالة أخرى إلى الوكيل. والذي ينتقل إلى الروبوت عبر الباعث. أخيرا، يتعين على الروبوت تنفيذ الإجراء الذي تم استلامه باستخدام مشغلاته. يتم تنفيذ الإجراء سالف الذكر بشكل تكراري حتى يحقق الروبوت هدفه أو لعدد معين من الفترات/الحلقات، أو أي حالة تحتاجها حالة الاستخدام. إذا يعمل الوكيل المسؤول عن اختيار الإجراء داخل متحكم المشرف ويتم الحصول على المشاهدات بواسطة متحكم الروبوت. بهذه الطريقة، يعمل إطار عمل deepbots كغلاف مما يعني أنه ينجز ويخفي عمليات معينة عن المستخدمين، حتى يتمكنوا من التركيز على مهمة DRL، بدلا من التعامل مع جميع التفاصيل الفنية الخاصة بالمحاكي. في الوقت نفسه، يعمل هذا الإطار أيضا على إثراء التدريب بميزات المراقبة الحية، والتي تساعد الباحثين في الحصول على مشاهدات حول الأجزاء الأساسية لعملية التدريب (Kirtas et al, 2020).



الشكل 4-5 تواصل المتحكم مع المشرف في إطار عمل DeepBots (Kirtas et al, 2020).

5.3.2.4. خوارزمية أمثلة السياسة القريبة Proximal Policy Optimization - PPO:

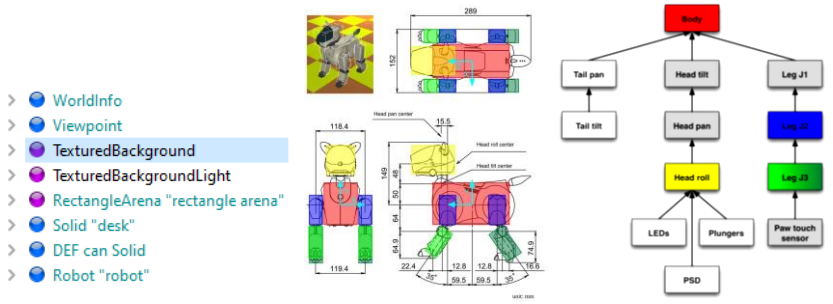
لا يزال تحديد أفضل خوارزمية لمهمة الإمساك محل نقاش العديد من الباحثين. رغم أن خوارزميات OFF-Policy and model free، هي الأكثر فعالية لتطبيقات التحكم المستمر مثل مهمة الإمساك. ورغم ذلك، لا يزال موضوع تجربة الخوارزميات الحديثة، ومقارنتها فيما بينها أمرا مهما في عالم التحكم الروبوتي (Sekkat et al, 2023). ومن بين الخوارزميات المختلفة برزت SAC وPPO لفعاليتهما في موازنة الاستكشاف والاستغلال، وهو أمر بالغ الأهمية للتعلم الفعال لسياسات التحكم (Shianifar et al, 2024). لذا، في هذه الرسالة استخدمت خوارزمية أمثلة السياسة القريبة PPO، وهي خوارزمية ON-Policy and model free، إذ ذكرت عدة أبحاث مدى قدرتها وتفوقها في العديد من المهام، فكما يوضح الشكل (5-5)، مقياس لأداء خوارزمية DQN وهي خوارزمية OFF-Policy and model free مع خوارزمية PPO في مهام روبوتية مختلفة وعلى عكس ما ذكر (Sekkat et al, 2023) نلاحظ تفوق واضح للخوارزمية PPO (Thompson et al, 2019). وبالمقارنة مع خوارزميات أخرى مثل A2C وA3C وTRPO وCEM، وجد أنها تتفوق على جميع الخوارزميات السابقة (Schulman et al, 2017). كما قدمت (Thompson et al, 2019) إحصائيات الأداء الكمي لخوارزميات RL (DQN, PPO, TRPO, Dyna-Q, MPC) في المهام الروبوتية كالملاحة والتلاعب والحركة تم التقييم من حيث المكافأة التراكمية ووقت التدريب حيث أظهرت خوارزمية PPO أداء لافتا.



الشكل 5-5 مقارنة DQN مع PPO من حيث الأنشطة الروبوتية المختلفة، مقابل المكافأة التراكمية (Thompson et al, 2019).

5.4. بناء بيئة المحاكاة:

هي المكان الذي سيتعلم ضمنه الوكيل ويقوم من خلالها بجمع بيانات التدريب بالقيام بإجراء ثم ملاحظة تأثيره على البيئة وقيمة المكافأة التي تم الحصول عليها من هذا الإجراء تحدد مدى أهميته مستقبلا بحيث يكرر أو لا يكرر وبالتالي هذه البيانات هي الأساس في تدريب الوكيل والحصول على سياسة اتخاذ قرار فعالة، لذا يجب بنائها بعناية واختيار خصائصها ومواصفاتها بدقة. يتم تمثيل البيئة في Webots كتمثيل شجري كما يوضح الشكل (5-6) كل فرع من هذه الشجرة يدعى (عقدة) لها اسم خاص بها ووصف وخصائص. داخل كل عقدة يوجد مجموعة أفرع ثانوية تسمى (أطفال العقدة) يكون لها أيضا اسم ووصف وخصائص محددة. وقد يكون لكل فرع أفرع ثانوية وهكذا... تتكون بيئة التدريب ضمن محاكي Webots، في هذا البحث من قسمين وهما:



الشكل 6-5 توضيح لمفهوم الشجرة والعقد وأطفال العقد في نموذج روبوت Aibo. وشجرة المشهد لبيئة التدريب ضمن Webots (Ijspeer et al, 2004).

1. العالم الجديد (WORLD): هي الأجزاء التي سيتعامل معها الوكيل أثناء تدربه ضمنها موضحة في الشكل (5-7) وهي تشمل كل أجزاء بيئة المحاكاة عدا الروبوت المدرب، مثل:
   - الجسم المراد إمساكه: وشكله يحدد القبضة التي سيقوم الطرف بالتدرب على إنشاؤها. يتم البدء بعلبة المشروبات المعدنية أولا لتوليد قبضة أسطوانية، ولاحقا لذلك سيتم التعديل حسب التجارب.
   - مكونات البيئة اللاحقة مثل (الأرضية، وطاولة، وإضاءة، وما إلى ذلك)، وشكلها وخواصها.

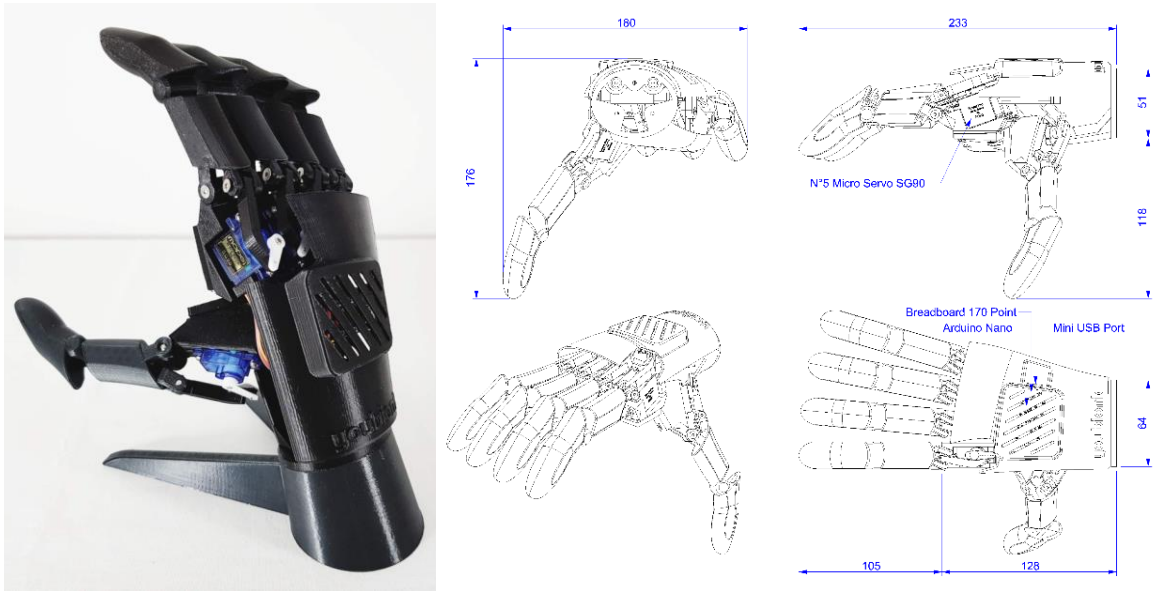
في هذه الخطوة تم إنشاء عالم جديد (HAND_DRL.wbt)، ثم ضبط إعداداته باستخدام العقدة worldInfo، التي تتحكم في خصائص البيئة الفيزيائية والمعرفات العامة لعالم المحاكاة. تم خلالها تفعيل الجاذبية $9.81\ m/s^2$، وضبط نوع محرك الفيزياء المستخدم وهو ODE المحرك الافتراضي في بيئة المحاكاة Webots، ثم عين زمن خطوة المحاكاة (basicTimeStep) (وهو الزمن بالميلي بين كل تحديث فيزيائي للمحاكاة)على قيمة 32ms وهي القيمة الافتراضية للبرنامج واعتمدتها معظم الدراسات السابقة تقليل هذه القيمة سيبطئ المحاكاة ولكن سيجعل النتائج أكثر دقة ورفع هذه القيمة سيجعل المحاكاة أسرع لكن سيؤثر على دقة النتائج يرتبط timestep بمحرك ODE لأنه يستخدم كفارق زمني Δt لحل معادلاته التفاضلية بالطرق التكاملية، تم تحديد optimalThreadCount عدد خطوط المحاكاة على قيمة 1، يمكن زيادتها إذا كان هنالك معالج متعدد الأنوية. ثم ضبط معامل الاحتكاك الانزلاقي والساكن بين الأجسام بناء على نموذج كولوم للاحتكاك coulombFriction الذي يحدد قوة مقاومة الانزلاق بين جسمين متلامسين على قيمة 1 وهي قيمة عادة تكون بين [0-2] وتمثل قيمة 1 قيمة افتراضية وسطية. تم تفعيل معامل التخميد defaultDamping الخطي والدوراني الذي يعمل كعامل فرملة ناعمة تبطئ الجسم تدريجيا عند تحركه على قيمة 0.2 لتحسن الواقعية والاستقرار الفيزيائي علما أن قيمة 0 تعني عدم التخميد نهائية وأي قيمة فوق 0.5 تعني تخميد قوي قد ينهي أي حركة بسرعة أما 0.2 قيمة افتراضية وسطية تخمد دون قتل الحركة. وبعد تعيين أهم الخصائص للبيئة تم الانتقال لتحديد عناصر البيئة المحيطة يتم إضافة عقدتين من نوع Solid أحدهما تمثل الطاولة التي سيتوضع عليها الأجسام والأخرى هي الجسم المراد التقاطه (علبة مشروبات معدنية ممتلئة)، وعقدتين آخرتين أرضية وإضاءة. (تستخدم تعليمات مثل translation وrotation لضبط الموقع والتوجيه للعناصر). كل هذه العناصر تستورد من مكاتب البرنامج وتكون علبة المشروبات المعدنية بوزن 340 غرام تقريبا.



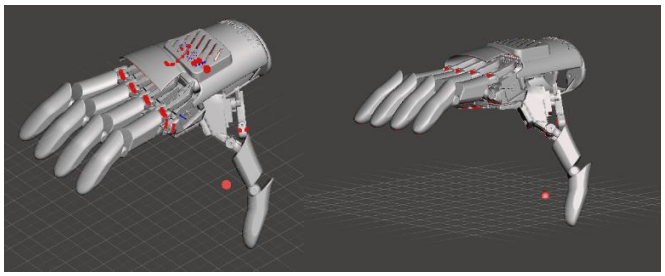
الشكل 7-5 بيئة التدريب.

2. نموذج الكف الاصطناعي المدرب (Robot Hand): وهو الكف الذي سيتم تدريبه على مهمة الإمساك. وبناء نموذج لليد المختارة ضمن برنامج المحاكاة المعتمد تم وفق مجموعة من الخطوات وهي كالتالي:

1) **<u>طلب نموذج</u>** يد HANDY من الشركة المصنعة وبسبب قيود عملية الدفع وبعض المشاكل في الحصول على الطرف الذي تقرر إجراء البحث عليه وفرت الشركة للبحث كملفات STL نموذج يد يدعى Handy Lite وهو نموذج يد يعتمد في الأساس على اليد HANDY وله نفس الأبعاد والخصائص ويزن نصف كيلو أيضا في أبعاده القياسية إلا أنه لا يملك محركات في المفاصل بين السلاميات البعيدة والقريبة ولكن يملك 6 محركات في المفاصل السلامية المشطية اثنين منها للإبهام قبل المحاكاة يتم تعديله بإضافة 5 محركات بين السلاميات البعيدة والقريبة ليصبح النموذج مع 11 درجة حرية فيعود ليمثل يد HANDY المختارة. يوضح أبعاد وشكل اليد من الشركة في الشكل (5-8). كما يوضح الشكل (5-9) اليد كملف STL قبل إجراء أي تعديل عليه. تشير درجات الحرية للنظام إلى عدد الوظائف التي يمكن تشغيلها بشكل مستقل. في الأنظمة الفيزيائية، درجات الحرية (Degree of freedom – DOF) هي عدد الحركات المستقلة التي يمكن للنظام أن يقوم بها. درجات التحكم (Degree of control - DOC) هي عدد الوظائف القابلة للتحكم والتي تتوفر فعليا للمستخدم، وقد تكون مجموعة فرعية من درجات حرية النظام أو مجموعات من درجات حرية متعددة (Sherstan, 2018).

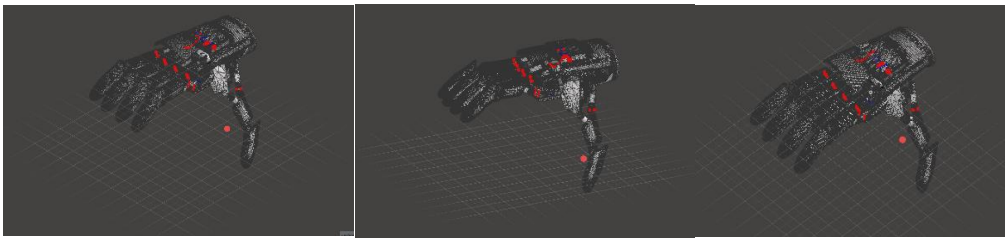


الشكل 8-5 نموذج يد Handy lite.

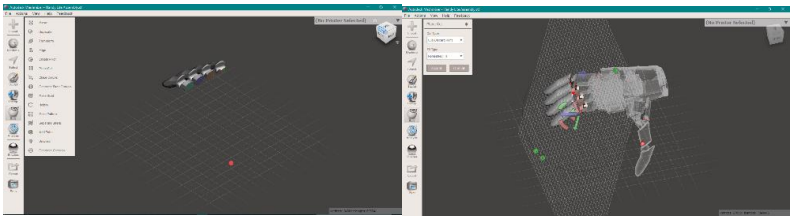


الشكل 5-9 صورة ليد Handy Lite ضمن برنامج MeshMixer كملف STL وفرته الشركة.

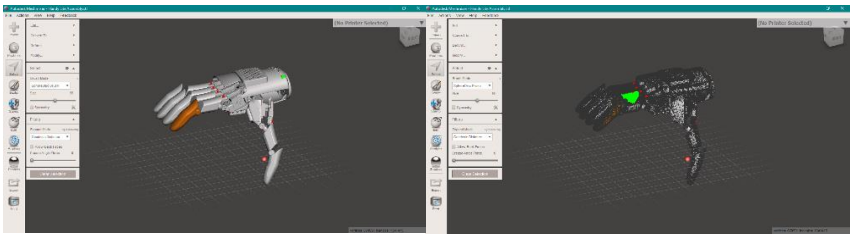
2) **<u>تهيئة نموذج اليد:</u>** إن الملف الذي وفرته الشركة يأتي كملف STL مجمع لكل أجزاء اليد فيشكل قطعة صلبة واحدة، واستخدامه هكذا ضمن بيئة Webots لن يمكن من إضافة محركات لذا تم تقطيع الشكل في أماكن عمل المحركات (المفاصل) إلى مجموعة من الأجزاء الصلبة وهي السلاميات وراحة الكف ليكون من الممكن لاحقا تجميع هذه الأجزاء من خلال مفاصل فيها محركات في برنامج Webots مما يتيح وجود حركة نسبية وبالتالي إمكانية التحكم في حركة المفاصل. تم تعديل النموذج باستخدام برنامج MESHMIXER (هو برنامج تعديل للنماذج ثلاثية الأبعاد يتيح تقطيع الأشكال، ودمجها، وحذف أجزاء منها، كما يتيح إمكانية تعديل الحجم للكتلة ثلاثية الأبعاد، وغيرها...)، يعرض نموذج اليد الأولي وهو ملف STL ضمن MESHMIXER كشبكة من العناصر المنتهية تم استخدام مجموعة تعليمات لتحديد أجزاء هذه الشبكة المرغوب تقسيمها كتعليمات plane cut، select – delete (الأكثر دقة)، ثم حفظ كل جزء من هذه اليد كقطعة صلبة منفردة بصيغة STL. تعرض الأشكال النموذج الصلب الأول لليد ثم عمليات التعديل وصولا إلى القطع النهائية لليد. إذ يعرض الشكل (5-10) شبكة العناصر المنتهية والتي نلاحظ كم هي دقيقة مما يدل على جودة الملفات لاحقا أثناء الطباعة، ولكنه أيضا قد يسبب مشاكل لاحقا في عملية المحاكاة بسبب التعقيد الحسابي. يوضح الشكل (5-11) عمليات التعديل مع تعليمة القطع المستوي Plan-cut يليها الشكل (5-12) ويعرض تعليمة تحديد- حذف – Select delete تكرر هذه العمليات للحصول على كل الأجزاء في الجدول (5-2).



الشكل 5-10 عرض لشبكة عناصر الجسم ضمن برنامج MeshMixer



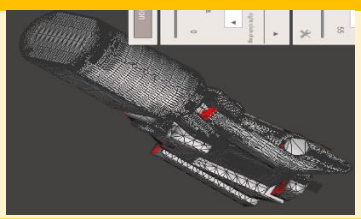
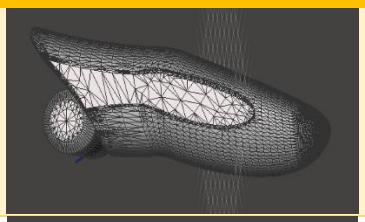
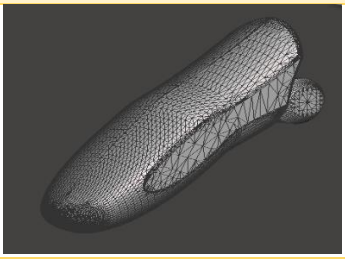
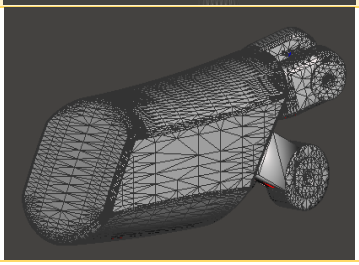
الشكل 5-11 استخدام تعليمات القطع المستوي Plane Cat.

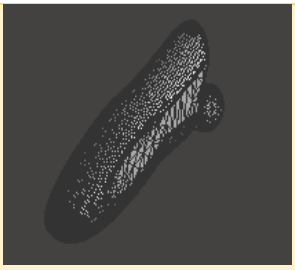
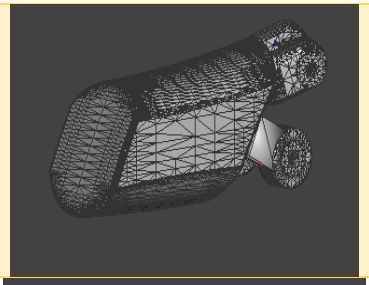
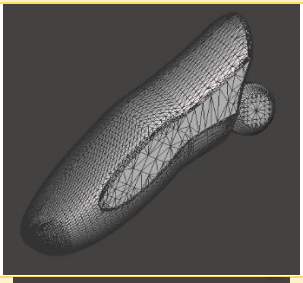
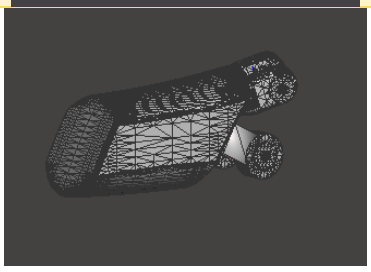
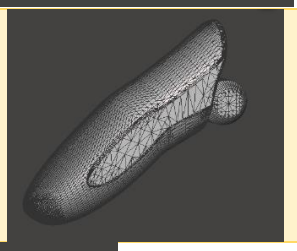
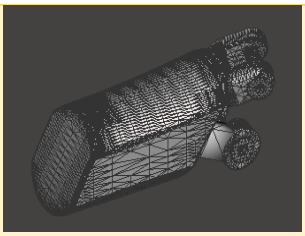
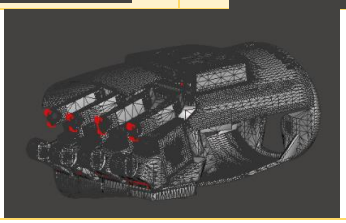


الشكل 5-12 استخدام تعليمة Select –Delete

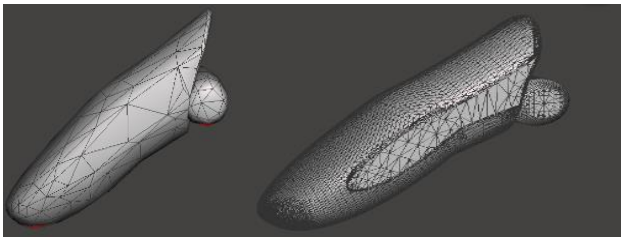
يعرض الجدول (5-2)، ناتج عمليات التقطيع وهي ملف راحة اليد وملفي الإبهام وملفي السبابة وملفي الوسطى وملفي الخنصر وملفي البنصر. وهي الملفات النهائية التي تم الحصول عليها، وحفظها بصيغة STL لاسترادها لاحقا في Webots أثناء عملية بناء الروبوت. كما يبين الجدول فإنها ذات شبكة مرتفعة الدقة بشكل كبير، وأثناء التجارب العميلة اللاحقة ضمن Webots شكّل هذا مشاكل في عملية التدريب بسبب تعقيدها والحمل الحاسوبي المرتفع الذي تتطلبه لإتمام المهمة. لذا تم تعديلها وإعادة تغير خصائص الشبكة وتقليل تعقيدها من خلال Reduce Mesh التي تستخدم لتقليل عدد العناصر المثلثية في الشبكة (Polygons/Triangles) في المجسم، أي تقليل تعقيد النموذج ثلاثي الأبعاد بالتالي تسهيل المعالجة والطباعة، وتقليل حجم الملف، وتحسين أداء البرنامج. تم اختيار نوع Shape Preserving عند تحديد Reduce type، وتعني هذه الخاصية التقليل مع الحفاظ على الشكل العام للمجسم. كما تم اختيار Max Deviation عند تحديد Reduce Target وهو ما يحدد كيفية تقليل التفاصيل، إذ حدد الحفاظ على أقرب شكل ممكن مع تقليل التفاصيل حسب حد أقصى للانحراف تم تعينه بالتجريب 0.5 مم. وحددت النسبة المئوية للتقليل على قيمة نصف (50%) في أغلب العناصر أو أكثر قليلا في بعض العناصر (30% في راحة اليد)، تم التقليل مع تفعيل خيار Preserve Boundaries الذي يحفظ حدود المجسم (الحواف الخارجية) من التشوه.

الجدول 5-2 ناتج عملية التقطيع والتهيئة.

الإصبع	السلاميات البعيدة	السلاميات القريبة
الإبهام		
السبابة		

الوسطى		
الخنصر		
البنصر		
راحة الكف		

يعرض الشكل (13–5) مثال قبل وبعد عملية تقليل الشبكة لسلامية البنصر، وكُررت هذه العملية لكافة الأجزاء. علما أنه حتى بعد التقليل لا تزال قيم العناصر في المكونات كبيرة مقارنة بأبحاث وتجارب أخرى (Kepes, 2017).



الشكل 13–5 السلامية البعيدة للبنصر، يمينا قبل تقليل الشبكة، يسارا بعد تقليل تعقيد الشبكة

3) **<u>الانتقال لبرنامج Webots:</u>** لبناء نموذج الكف ضمن المحاكي نحتاج لتحديد عدد من المحددات وهي:

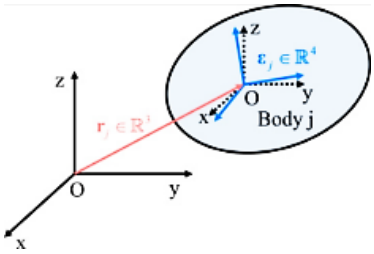
- الهيكل الخارجي وهذا تم تجهيزه في الخطوة السابقة.
- أدوات التشغيل وموضع المحركات في المفاصل.
- أدوات التغذية الراجعة (المستشعرات).
- المواد وخصائصها ومعاملات الاحتكاك.
- حدود الزوايا للمفاصل.

بناء هيكل الروبوت: كما ذكر سابقا تمثل Webots البيئة كمشاهد ذات بنية شجرية تكون فيها العقدة الجذرية هي العالم والعقد الفرعية الخاصة بها هي العناصر المختلفة في العالم. وبالتالي، يجب أن يكون الروبوت عقدة تحت عقدة الجذر تحتوي على وحدة تحكم. **عقدة الروبوت** هي تمثيل لجسم فيه محركات وحساسات وكتل صلبة لكل منها خصائص فيزيائية وأشكال محددة ويمكن التحكم به. تم تجهيز الهيكل الخارجي كقطع حفظت بصيغة STL في الفقرات السابقة وسيتم استخدام هذه الملفات لبناء نموذج اليد. إذ يمكن لـ Webots استيراد نماذج ثلاثية الأبعاد من معظم برامج النمذجة ثلاثية الأبعاد من خلال معيار VRML97. تم استيراد كل جزء بصيغة STL. يكون للروبوت وحدة تحكم من أجل قراءة القيم من أجهزة الاستشعار والتصرف وفقا لذلك، وهي نصوص برمجية مسؤولة عن وظائف العقدة. لأغراض RL وDeepBots، من الضروري تضمين عقدة مشرف خاصة لديها معرفة كاملة بالعالم ويمكنها الحصول على معلومات وتعديل كل العقد الأخرى. على سبيل المثال، يمكن استخدام المشرف لنقل العناصر في العالم أو قياس المسافات بين الروبوت والهدف. إذا لبناء نموذج اليد يتم إضافة عقدة من نوع ROBOT، تضم هذه العقدة عدة عقد فرعية تسمى Children والتي تضم أجزاء اليد، بالإضافة للعقد الأبناء تملك العقدة الأم، مجموعة من الحقول وعدة خصائص لضبط العوامل والكتل المستخدمة في الروبوت، أهمها:

- الأبناء (Children): وهي القطع المكونة للروبوت، في البحث هي (قاعدة التثبيت والتي تحتوي مفاصل ثابتة، السلاميات وراحة الكف، المحركات والحساسات)، تعرض آلية ربطها وخصائصها لاحقا.
- BoundingObject: هو الشكل الهندسي الذي يستخدمه محرك ODE لحساب التصادمات يجب أن يأخذ شكل الروبوت والأجسام. بدونه لن يتفاعل الجسم أو القطعة مع الأجسام الأخرى.
- Physics: تحتوي على إعدادات الكتلة، مركز الكتلة، عزم القصور الذاتي. بدونها يعتبر الجسم بلا وزن أو لا يتأثر بالجاذبية حيث يحتوي حقل mass.
- Supervisor controller: هو عقدة برمجية خاصة تسمح بالتحكم الكامل بالمحاكاة، يمكنه مراقبة أو تحريك الروبوتات أو تحريك الأجسام وتعديل الخصائص للكائنات، ويستخدم للتحكم في المشهد من خارج الروبوت.

بالنسبة للعقد الأبناء ضمن عقدة الروبوت فهي من عدة أنواع وأشكال تربط فيما بينها بطريقة منظمة وتحدد خصائصها بدقة في البداية يتألف كامل الهيكل من عقد Solid (أجسام فيزيائية حقيقية) وهي الملفات الناتجة عن عملية التقطيع يمكنها تمثيل الاصطدام والتحرك والدوران وما إلى ذلك وهي راحة اليد والسلاميات داخل كل عقدة solid يوجد فيزياء يحدد ضمنها كثافة المادة اعتبرت مادة الروبوت PLA التي تملك كثافة حجمية $1.24\ g\backslash cm^3$ (Barrasa et al, 2021). فيكون الطرف ككل نصف كيلو غرام تقريبا وهي قيمة قريبة جدا من الواقع ضمن الروبوت يتم إضافة هذه الكثافة لكل سلامية ولراحة اليد يعمل ذلك على جعل مركز ثقل اليد أقرب ما يمكن للواقع. بالإضافة للفيزياء يوجد حدود وأبناء (غالبا ما تكون عقد نوع shape لإظهار شكل الجسم الصلب). ولربط هذه الأجزاء فيما بينها وتحريكها بشكل صحيح نسبة لمبدأ الإحداثيات العالمي فاستخدمت العقدة Transform –لتحديد إطار (موقع

واتجاه) الجسم المضاف نسبة للفضاء الإحداثي ثلاثي الأبعاد العالمي. فيتم تحريك قطع الإطار (الذي يتضمن أجسام وأشكال داخله) نسبة للإطار العام وتوَضّح الفكرة في الشكل (14–5) حيث يمثل BODY J إطارا مرجعي ثانوي لدراسة حركة واتجاه الأجسام ضمنه نسبة للفضاء العالمي– بالإضافة لعقد أخرى كالمفاصل ومحركات المفاصل ومستشعرات اللمس والزاوية للمحرك وما إلى ذلك ستذكر مع خصائصها لاحقا. تم تثبيت الروبوت على جسم صلب لا يتحرك لأن هذا البحث يركز على قضية الإمساك دون الانتقال لأن مرتدي الطرف يقوم بنفسه بتوجيه الكف فلا حاجة لهذه المهمة. يتم توضيح النموذج في الشكل (15–5).



الشكل 14–5 مفهوم تحويل الإطار بين جملتين إحداثيتين Transform Frames (Luo et al, 2023).



الشكل 15–5 النموذج الذي تم بناءه قبل إضافة المحركات والحساسات.

مواضع المفاصل وأدوات التشغيل (المحركات) ضمنها: للكف أحد عشر مفصلا في الأصابع، في كل إصبع مفصلين (محركين) عدا الإبهام فيه ثلاث محركات اثنين منهم في قاعدته لتوليد حركتين أحدهم لتوليد حركة (تقريب تبعيد) والثاني (أنثي وحشي)، في هذا البحث تمت نمذجة كافة المحركات الإحدى عشرة التي تعطي النظام إحدى عشرة درجة حرية (11 DOF) لكن أثناء عملية التحكم تم إلغاء حركة المفصل (أنثي وحشي) في قاعدة الإبهام بالتالي يكون هناك عشرة درجات تحكم (10 DOC)، يوضح الشكل (16–5) أماكن توزع المحركات الإحدى عشر. كما يعرض الجدول (3–5) المفاصل ومواقعها ودرجات الحرية فيها ومجال الحركة لكل إصبع.

الجدول 3–5 عدد المفاصل ومواصفاتها التي تمت نمذجتها، (اللون البرتقالي لن يتم التحكم به لاحقا).

الإصبع	DOF	عدد المفاصل	اسم المفاصل، نوع الحركة ومجال دورانها
الإبهام	3	2	MCP 2 - DOF ABD\ADD [-10° - 57°] FLEX\EXTE [-57° - 57°] PIP 1 – DOF - FLEX\EXTE [[0° - 115°]]
السبابة	2	2	MCP 1-DOF - FLEX\EXTE [0° - 115°] PIP 1 -DOF - FLEX\EXTE [0° - 85°]
الوسطى	2	2	MCP 1-DOF - FLEX\EXTE [0° - 115°] PIP 1-DOF - FLEX\EXTE [0° - 85°]
الخنصر	2	2	MCP 1-DOF - FLEX\EXTE [0° - 115°] PIP 1-DOF - FLEX\EXTE [0° - 85°]
البنصر	2	2	MCP 1-DOF - FLEX\EXTE [0° - 115°] PIP 1-DOF - FLEX\EXTE [0° - 85°]

Metacarpophalangeal Joint – MCP مفصل سنعي سلامي.
Proximal Interphalangeal Joint -PIP مفصل بين السلاميات القريبة.

نوع المحرك في هذه المفاصل هو MicroServoSG90 يمكنه الدوران °180 بعزم دوران 2.5 Kg-cm. يتم التحكم فيه من خلال الدرجات إذ يمكنه أن يدور بسرعة تشغيل 0.1/60s (SERVO MOTOR SG90 DATA SHEET). عادة تكون أصغر درجة للتحكم فيه °1 وقد يسمح بحالات خاصة بزوايا أكثر دقة في هذا البحث سيتم ضبط °1 كأصغر زاوية يمكن أن يدورها المحرك. في Webots يتم تمثيل هذه المفاصل مع محركاتها من خلال عقدة مفصل HingeJoint وهي مفاصل تسمح بالحركة الدورانية حول محور واحد فقط تربط الأجزاء Solid فيما بينها ولكل مفصل خصائص:

- JOINTPARAMETERS: لها مجموعة خصائص واجب ضبطها، وأكثرها أهمية axis التي تحدد محور الدوران، وanchor التي تحدد موقع نقطة الدوران، ويجب أن تضبط بدقة للحصول على الحركة الصحيحة.
- JOINTDEVICE: يوضع ضمنها المحرك وحساسات زوايا المحرك. في هذا البحثي هي:
  - PositionSensor: يعيد هذا الحساس موضع المفصل بالراديان.
  - RotationalMotor: يتم استخدام محرك دوراني وضبطه برمجيا على نمط position control mode ليتم التحكم فيه من خلال الزوايا لمحاكاة servo motor المتحكم بها من خلال الزاوية، كما يتم ضبطه من خلال عدة خصائص كالسرعة والعزم والموضع والتسارع وتم ضبطه في هذه الدراسة على خصائص السيرفو التي ذكرت في الأعلى (SERVO MOTOR SG90 DATA SHEET).
  - ENDPOINT: يربط خلالها الجزء التالي والذي سيتحرك مع المفصل ويكون نوع SOLID غالبا.



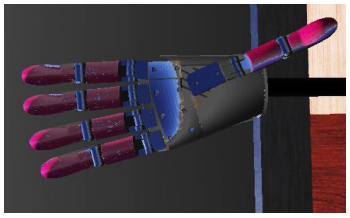
الشكل 16-5 يوضح الشكل جمع أجزاء الهيكل مرتبطة ببعضها من خلال المفاصل (باللون الأحمر).

تحديد مواضع التغذية الراجعة (حساسات اللمس): في هذا البحث يتم تعديل الكف النظامي بإضافة حساسات لمس، ستستخدم مدخلاتها لاحقا للتحكم في زوايا المفاصل. يوجد العديد من الأنواع المتاحة تجاريا، عادة في الأطراف الاصطناعية بعد اختيار نمط الحساس يتم وضعه على الأماكن المرغوبة على اليد مع طبقة من السيليكون، في هذا البحث سيتم اعتماد 10 حساسات موزعة على السلاميات ورؤوس الأصابع. يوضح الشكل (17-5) توزيع هذه الحساسات. الاستشعار اللمسي (Tactile Sensing) المستخدم في هذه الدراسة يتضمن نوعين:

- استشعار لمسي ثنائي (Binary Touch Sensing): يُستخدم للكشف عن وجود أو غياب التلامس بين الأصابع والهدف، ويُعبَّر عنه بإشارة رقمية (0 أو 1).
- استشعار قوة التلامس (Contact Force Sensing): يُمكِّن من قياس شدة القوة المطبقة عند نقطة التلامس، ويوفر بيانات كمية مستمرة تعكس التفاعل الفيزيائي بين اليد والجسم.

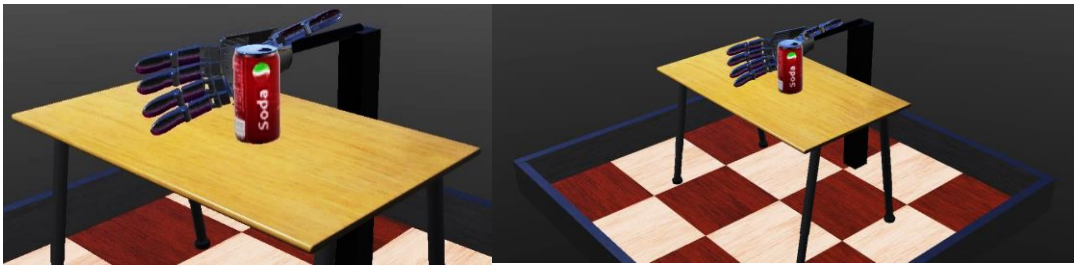
في Webots تُستخدم عقدة TouchSensor لنمذجة مستشعرات اللمس. يعتمد اكتشاف التلامس على حقل "boundingObject" لعقدة TouchSensor وحقل "boundingObject" للعقد الصلبة الأخرى التي يلامسها. يتم نمذجة مستشعرات الروبوت الخاص بالدراسة هذه من خلال تعيين حقل "type" الخاص بها على:

- حالة الدراسة الأولى:"bumper" والتي تعيد قيما ثنائية (1\0)، تكشف عن وجود التلامس من عدمه.
- حالة الدراسة الثانية:"force" والتي تعيد مقدار القوة المطبقة بين الجسمين بالنيوتن.

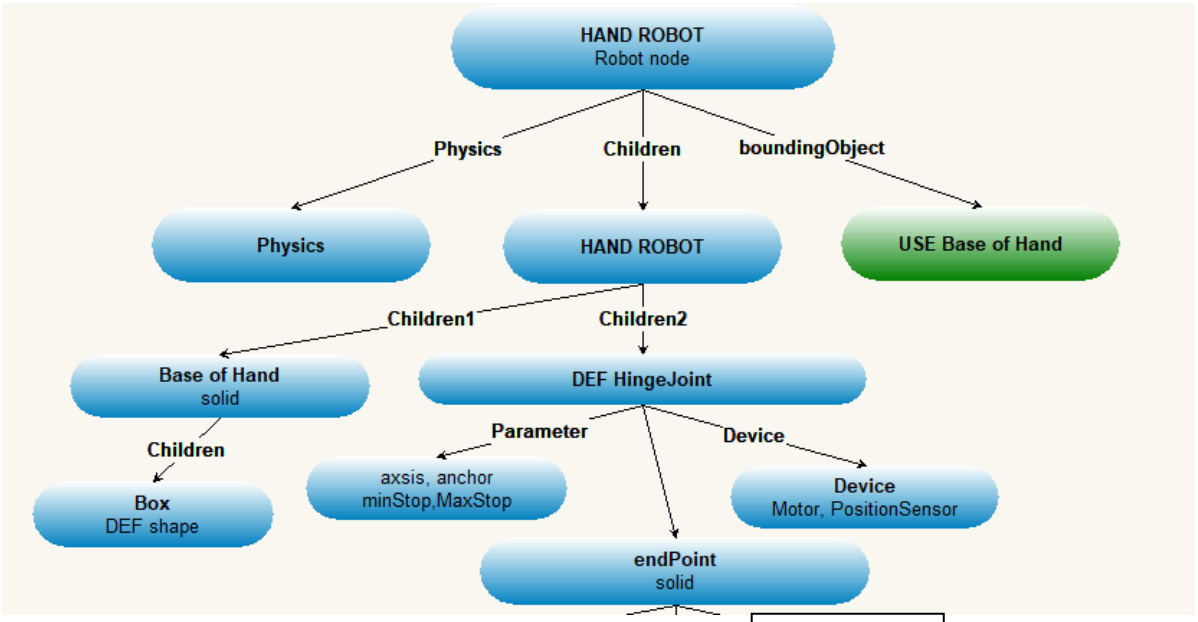


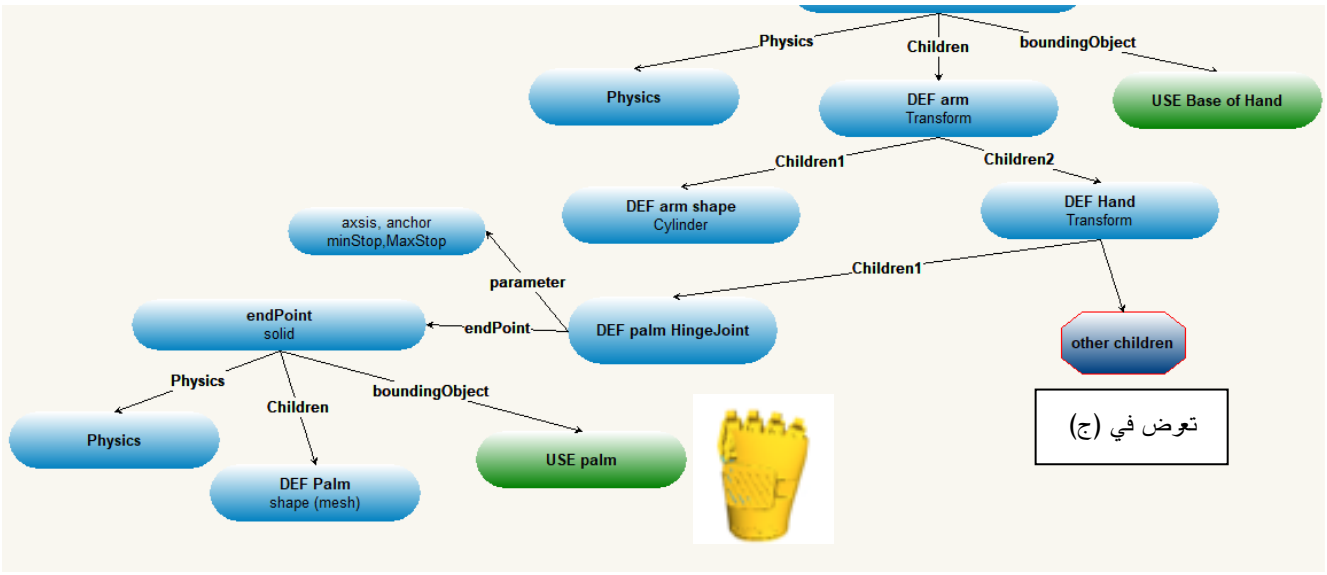
الشكل 17-5 توزع الحساسات على الكف.

تصبح بيئة المحاكاة بالكامل الناتجة عن جمع المكونين السابقين كما في الشكل (18-5). كما ويوضح الشكل (19-5) العلاقات الهرمية بين جميع عقد النموذج التي ذكرت، بما في ذلك تلك التي لم يتم ذكرها بعد. حيث يتم استخدام هذه البلوكات كما هو معروض، لتكوين اليد الروبوتية المثبتة على ذراع عام ثابت بمفصل معدوم الحركة.

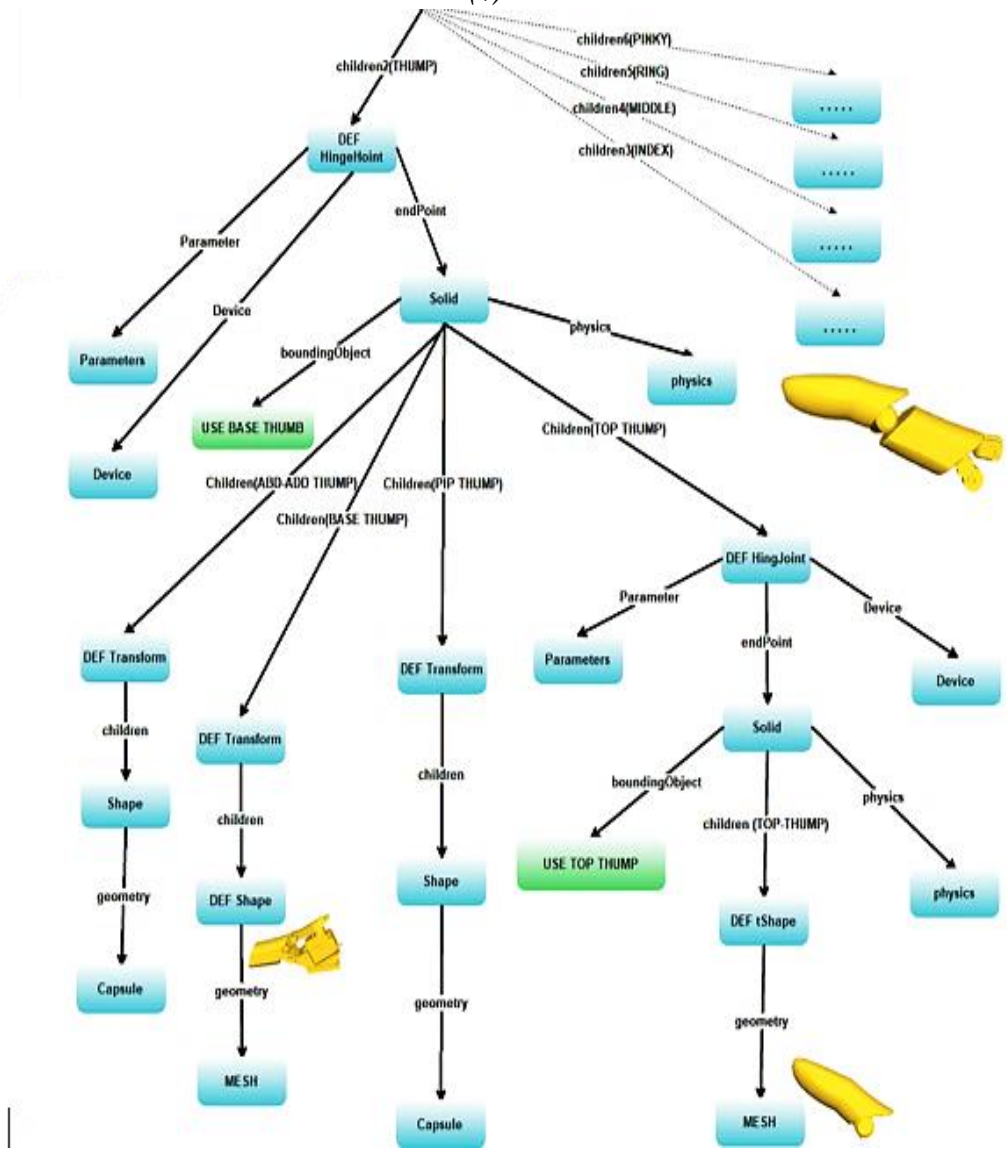


الشكل 18-5 بيئة التدريب المبنية ضمن برنامج Webots.





(ب)



(ج)

الشكل 19–5 العلاقات الهرمية بين جميع عقد النموذج (أ وب) يمثل هيكل التثبيت الخارجي وراحة الكف و(ج) يمثل الأصابع وربطها براحة الكف. ملحوظة: تعرف معظم العقد بالتعريف DEF يعني أنه يمكن إعادة استخدام خصائصها، التعريف USE يعني استخدام عقدة معرفة مسبقا.

5.5.حالتي الدراسة (Case Studies):

في هذا القسم، سيتم عرض حالتين دراسيتين تهدفان إلى تطوير نظام تحكم قائم على التعلم المعزز العميق لتدريب اليد الروبوتية متعددة الأصابع ضمن بيئة Webots، لإيجاد سياسة التحكم الأمثل لتوليد قبضة أسطوانية تحت ظروف استشعار مختلفة. الحالة الأولى تعتمد على دمج حساسات لمس ثنائية تحدد وجود التلامس بين الكف والجسم الملتقط لتغذية النظام بمعلومات التلامس، بينما الحالة الثانية تستخدم حساسات قوة اللمس لتوفير إدراك أكثر واقعية حول التفاعل مع البيئة. تهدف عملية تحليل فعالية نموذج DRL في التعلم والتحكم تحت نوعين مختلفين من المدخلات الحسية، إلى تحديد مدى تأثير نوع البيانات المستقبلة على استقرار وكفاءة الأداء.

5.5.1.حالة حساسات اللمس:

في هذه الحالة الدراسية، تم الاعتماد على حساسات لمس لتوفير بيانات حول تلامس اليد الروبوتية مع البيئة المحيطة، وذلك بهدف تحسين دقة التحكم والاستجابة في التفاعل الفيزيائي. تم تصميم بيئة التعلم بحيث تعتمد السياسة المتبعة على الإشارات اللمسية كمدخلات رئيسية لاتخاذ القرار، وتمت هذه المرحلة كالتالي:

5.5.1.1. صياغة المشكلة وفق عملية قرار ماركوف:

كما ذكر في الفقرة (4.3.3) يجب صياغة مشاكل RL وفق إطار عملية قرار ماركوف حيث يُمثل الوكيل البرمجي كعنصر يتفاعل مع البيئة بشكل متسلسل من خلال اتخاذ إجراءات بناءً على ملاحظاته بهدف الحصول على مكافآت تعزز سلوكه الناجح. ويسعى الوكيل لتعظيم المكافأة التراكمية مع مرور الزمن. في هذه الحالة تم بناء نظام DRL وفق نموذج MDP باستخدام حساسات لمس كمصدر للمعلومات ويتكوّن هذا النظام من العناصر التالية:

1. **الوكيل الذكي:** يمثل الكيان المتعلم أو متخذ القرار. ويقوم في كل خطوة زمنية بـ: تلقي مشاهدة من البيئة، وتنفيذ إجراء معين وتلقي قيمة عددية للمكافأة. وفي البحث تم تطوير الوكيل باستخدام خوارزمية PPO.
2. **البيئة:** هي العالم الخارجي الذي يتفاعل معه الوكيل الذكي وتشمل: استقبال إجراء من الوكيل الذكي، وتحديث حالتها استجابة للإجراء، وإرسال قيمة المكافأة للوكيل، والمشاهدة التالية للوكيل. تم بناء بيئة المحاكاة في الفقرة (5.4)، وهي بشكل عام طاولة تحتوي جسما أسطوانيا (علبة مشروبات معدنية) موضوعا أمام اليد الروبوتية. تم تعيين زمن خطوة المحاكاة بـ Simulation Time step= 32 ms.
3. **الحالة المرصودة:** تتكون من متغيرات تصف البيئة. في البحث تم استخدام (زوايا المفاصل العشرة المتحكم بها وكذلك قيم حساسات اللمس العشرة المثبتة على أصابع اليد وهي في هذه الحالة من البحث قيمة ثنائية [0,1] حسب إذا وجد تلامس أم لا، حيث يشير الرقم 1 لحالة التلامس والتفعيل، وأما الرقم 0 يدل على عدم التلامس وعدم التفعيل)، بالتالي، تتكون الحالة المدخلة من عشرين متغيرا في كل لحظة.

Observation Space (20)
10 from Touch Sensors (Binary) 10 from Joints Positions Sensors (Continuous)

4. **الإجراء:** يمثل التغيير الذي يحدثه الوكيل على البيئة. وقد تم دراسة حالتين:

A. الإجراء **A** لكل مفصل: هو إما زيادة زاوية المفصل بمقدار **C** أو تقليلها بـ **C** أو تركها دون تغير.

B. الإجراء **B** لكل مفصل: هو إما زيادة زاوية المفصل بمقدار **C** أو الإبقاء عليها ثابتة.

Action Space A Discrete (3)
Increase Angle: angle (joint_i, t+1) = angle (joint_i, t) + c Decrease Angle: angle (joint_i, t+1) = angle (joint_i, t) – c Leave angle unchanged: angle (joint_i, t+1) = angle (joint_i, t)
Action Space B Discrete (2)
Increase Angle: angle (joint_i, t+1) = angle (joint_i, t) + c Leave angle unchanged: angle (joint_i, t+1) = angle (joint_i, t)

5. **المكافأة:** قيمة عددية يتلقاها الوكيل بعد تنفيذ الإجراء، وتستخدم لتوجيه التعلم نحو السلوك المطلوب. أحد أكثر الأجزاء صعوبة هو تصميم تابع مكافأة بأكبر قدر ممكن من الكفاءة، نظرا لكونه رد الفعل الوحيد الموجِّه للوكيل لمعرفة ما إذا كان يقوم بإجراءات صحيحة أم لا، مثلا يمكن المكافأة على سبيل المثال على تحقيق اللمس، أو المعاقبة عند عدم تفعيل أي حساس، وهكذا.. يختلف التابع كثيرا تبعا لهيكلية الملاحظات والإجراءات ونوع المهمة الموكلة للوكيل، في هذه الحالة من البحث، تم تجريب هيكلين:

A. هيكل المكافأة **A**: يمنح مكافأة ثابتة في كل خطوة تضاف لمكافأة الحلقة إذا تحقق الشرط التالي (حساس اللمس في أعلى الإبهام وأي حساسين آخرين من باقي الأصابع مفعلة).

B. هيكل المكافأة **B**: يعتمد على النسبة المئوية لعدد حساسات اللمس المفعلة.

وبالتالي هي من النوع (الكثيف) إذ تقدم في كل خطوة وليس في نهاية الحلقة إذا تمت المهمة فقط. وتم تقييس المكافآت بعد الحصول عليها ضمن المجال **[1−0]**، للحصول على تدريب مستقر لاحقا، وتقليل حساسية السياسة للضوضاء، والحفاظ على تناسق واحد للقياس يمكّن من عملية قياس التقارب والمقارنة لاحقا.

Reward Function A Dense Reward

if (top touch sensor enable) and (any 2 other touch sensor enable) then:
Reward = reward + 10
else:
Reward = reward

Reward Function B Dense Reward
Percentage of Enable Touch Sensor Rescale [0-1]

6. **سياسة اتخاذ القرار:** تمثل الإستراتيجية التي يستخدمها الوكيل لاختيار الإجراءات بناء على الحالة. وسيتم تدريب الوكيل باستخدام خوارزمية PPO والتي تستخدم سياسة خالية من النماذج يتم تدريبها بشكل مباشر أي (Model-Free, ON-Policy). وأما دالة الانتقال، التي تعبر عن احتمال الانتقال من الحالة S إلى الحالة S´ عند تنفيذ الإجراء A. ونظرا لاستخدام خوارزمية Model-Free فإن السياسة لن تعتمد على نموذج صريح لدالة الانتقال بل يتم تعلمها مباشرة من خلال التفاعل مع البيئة في Webots. دون الحاجة

لتعريف نموذج صريح وهذه هي ميزة Model-Free RL مقارنة بـ Model-Based. تم استخدام PPO بأسلوب الفاعل – الناقد. عيّنت المحددات النهائية لها من خلال اقتراحات DEEPBOTS ومن خلال الضبط التجريبي وفق مبدأ التجربة والخطأ والتوصيات الشائعة في الأدبيات، وستعرض لاحقا.

5.5.1.2. أساليب التقييم:

قبل بدأ سلسلة التجارب لاعتماد السياسة الأنسب ولاحقا اختبارها لا بد من اختيار مقاييس تقييم تعكس الأداء بشكل صحيح لفهم مدى جودة الأداء وما إذا كان هناك أي داع للتحسين أو إنهاء التدريب والتثبيت. في RL يوجد العديد من المقاييس تم اختيار مجموعة بحيث تعكس الأداء وتمكن من المقارنة مع الدراسات السابقة في حالة اللمس هي:

- **مقاييس الكمية:**
  - المكافأة التراكمية Cumulative Reward - CR: هي مجموع المكافآت التي حصل عليها الوكيل على مدار عدة خطوات أو حلقة كاملة، وتشير إلى النجاح الإجمالي للوكيل في تحقيق هدفه. وتعني المكافأة التراكمية الأعلى أداء أفضل، لذا يجب على الوكيل أثناء التدريب العمل على زيادتها. سيتم ترميزها في التجارب EP_SCR. وتعين قيمتها الحدية التي تعني نجاح الحلقة (IS-DONE) بالتجريب.
  - متوسط المكافأة التراكمية على مدار عدة حلقات ويدعى بالمتوسط المتحرك Simple Moving Average - SMA، يقدم نظرة عامة عن الاتجاه العام لنتائج قيم المكافآت التراكمية. ويحدد مدى ثبات الأداء عبر الحلقات المختلفة.
  - وقت التقارب Policy Convergence: وهو الزمن أو عدد الحلقات التي استغرقها الوكيل للوصول إلى مستوى ثابت نوعا ما أو قريب جدا من النتائج المرغوبة.
  - معدل النجاح Success rate: وهي النسبة المئوية للحلقات التي يصل فيها الوكيل إلى حد النجاح ونسبة إكمال الوكيل لمهمته، وتحسب بالعلاقة:

$$\text{Average \% Grasp Success} = \frac{N_successful\ grasp}{Total\ number\ of\ trials} * 100\% \qquad (5\text{-}1)$$

  - صلابة السياسة: هي القدرة على الحفاظ على أداء ثابت في ظل وجود تغيرات أو اضطرابات في البيئة. ويشير لمدى قدرة الوكيل على التعامل مع التقلبات والمواقف الجديدة.
- **مقاييس نوعية:** تعكس هذه المقاييس أداء السياسة بطريقة مرئية من خلال مشاهدة أداء السياسة ضمن المحاكاة، وهي ثانوية يجب أن تدعم المقاييس الكمية ولا يعتمد عليها بشكل فردي.

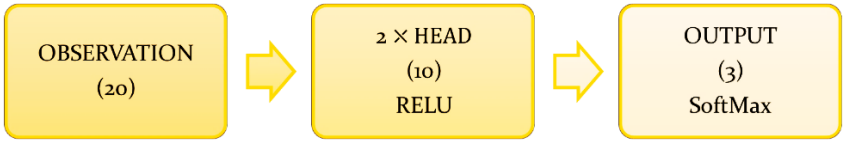
5.5.1.3. بناء وتدريب سياسة اتخاذ القرار:

بعد تعريف بنية المشكلة وفق نموذج عملية قرار ماركوف، تم الانتقال إلى مرحلة تطوير وتدريب سياسة اتخاذ القرار الخاصة بالوكيل البرمجي، وذلك باستخدام خوارزمية PPO. تهدف هذه السياسة إلى تعلم استراتيجية مثلى تمكن

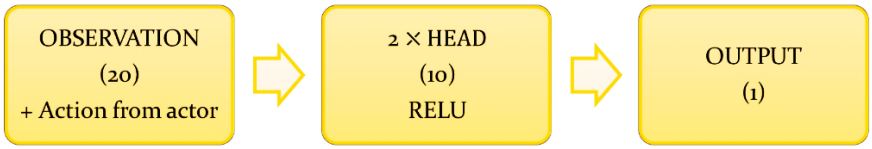
الوكيل من تحقيق التلامس المطلوب استنادا إلى المعلومات الحسية الواردة من حساسات اللمس. تم تنفيذ سلسلة من التجارب لاختبار فعالية عدة إعدادات، شملت التنويع في تعريف تابع المكافأة، وبنية الإجراء، وإعدادات الخوارزمية ومحدداتها، بهدف دراسة تأثيرها على جودة التعلم وسرعته.

1. تعيين بنية شبكة الفاعل – الناقد والمحددات الفائقة للنموذج المدرب:

تعد المحددات الفائقة محددات تصف النموذج المدرب، ولكن لا يتعلمها بذاته. مثل عدد الطبقات، ووظائف التنشيط، وعامل الخصم، وغيرها. وأحد أوجه القصور في معظم خوارزميات تعلم الآلة في الوقت الحالي هو اعتمادها على عدد كبير من المحددات الفائقة، والتي إن لم تُعيّن بعناية يمكن أن تجعل النموذج يفشل في معرفة الوظيفة الأساسية للبيانات. بالتالي يمكن أن يؤثر اختيار المحددات الفائقة بشكل كبير على منحنى التعلم بالإضافة إلى الأداء النهائي لسياسة اتخاذ القرار الناتجة. وعليه فإن هشاشة DRL للمحددات الفائقة تعني أن تحسينها، له أهمية كبيرة ويجب أن يتم إجراؤه لكل بيئة. يتم اختيار مجموعة قابلة للتطبيق من المحددات الفائقة للتعلم المعزز العميق من خلال إجراء عدد من التجارب المختلفة تُستخدم للبحث بشكل متكرر في مساحة المحددات الفائقة للعثور على مجموعة توفر أفضل النتائج وفقا لبعض المقاييس، ومن حيث التعلم المعزز تعد هذه المقاييس مكافأة يتمكن العميل من تجميعها على مدار فترة تقييم معينة كما نعلم. في هذه الحالة من البحث تم إجراء التحسين اليدوي للمحددات الفائقة (علما أن الطرق الآلية باستخدام إطار Optuna مثلا قد تكون نافعة) بهدف الحصول على مجموعة من المحددات الفائقة التي تسمح بالتعلم القوي للسياسة ضمن البيئة التي تم إنشاؤها والملاحظات وخوارزميات التعلم المعزز المستخدمة. والمحددات المستخدمة في هذه التجارب تعرض في الشكل (20–5) والجدول (4–5) وهي: في بنية الفاعل الناقد يوجد شبكتين عصبيتين، هما شبكة الفاعل، تتعلم السياسة. وشبكة الناقد، تتعلم تقدير دالة القيمة لتقيم الفاعل وتصرفاته. الشبكتان مصممتان في هذا البحث كبنى بسيطة تحتوي كل منهما على طبقتين مخفيتين مكونتين من 10 عقد لكل طبقة، مع استخدام تابع التفعيل ReLU بعد كل طبقة. شبكة الفاعل تنتهي بطبقة إخراج مع تابع Softmax لتوليد توزيع احتمالي للأفعال، بينما تنتهي شبكة الناقد بطبقة وحيدة العصبون دون تابع تفعيل لإخراج تقدير قيمة الحالة. تم استخدام محسن Adam المعروف لكل من شبكتي الفاعل والناقد، مع تحديد معدل تعلم قدره 0.001 للفاعل و0.003 للناقد. وهي القيم الافتراضية التي يقدمها DEEPBOTS عند استدعاء PPO والتي تحقق التوازن بين الاستكشاف والاستغلال، ويكون هذا التوازن أكثر معامل مؤثر على نجاح التعلم (Shianifar et al, 2024). يتم حفظ الحركات في ذاكرة انتقال مؤقتة (Replay Buffer)، ومن ثم تستخدم لتحديث النموذج بطريقة مصغّرة باستخدام مقاطع فرعية عشوائية من البيانات المجمعة (Mini-batch updates).



(أ)



(ب)

الشكل 20–5 بنية شبكتي الفاعل (أ) والناقد (ب).

الجدول 4–5 المحددات الفائقة لخوارزمية PPO.

المعامل	القيمة	الشرح
ε Clip-param	0.2	حد القطع للنسبة بين السياسات
max_grad_norm	0.5	أقصى قيمة لمقدار التدرج المسموح بها ويمكن استخدامه لقص التدرجات أثناء الأمثلة لمنع "انفجار" التدرجات
PPO_update_iters	5	عدد تكرارات تحديث PPO في كل دورة
Batch -Size	8	حجم العينة المستخدمة في التحديث
γ Gamma	0.99	عامل الخصم لتقدير المكافأة التراكمية المستقبلية عادة يكون ضمن المجال [0-1]، وبالتالي يحدد أهمية المكافأة المستقبلية والقيم القريبة من 1 يدل على إعطاء أولوية للمكافآت المستقبلية.

2. اختبارات تدريب السياسة (اختبارات المرحلة الأولى):

في هذه المرحلة يتم إجراء سلسلة من التجارب بهدف تعين تابع المكافأة الأنسب ومساحة الإجراءات التي تعطي أنسب سياسة، فيها تم تدريب الوكيل على تنفيذ مهمة الإمساك بعلبة المشروبات المعدنية ضمن بيئة Webots. في البداية نفذت تجارب مكثفة تقوم على مبدأ التجربة والخطأ لتحديد السياسة المثلى، وأهم هذه التجارب هي:

1) أولا، (الاختبار الأول) تم استخدام الإجراء A مع هيكل المكافأة A. وذلك ضمن 5000 حلقة تدريب تضم كل منها 700 خطوة وبذلك يكون للوكيل 3.5 مليون خطوة تدريب، لقياس تقدم التدريب يلزم مقياس يعكس مدى نجاح الضبط الحالي بشكل صحيح، والأكثر وضوحا هو تتبع إجمالي المكافأة المستلمة لكل حلقة (المكافأة التراكمية) التي تظهر نجاح الوكيل وسرعة التدريب وإكمال المهمة، لذا تم تقييم عملية التدريب من خلال متابعة قيمة المكافأة التراكمية اعتبرت المهمة ناجحة في الحلقة الواحدة إذا استطاع الوكيل الذكي الوصول لقيمة مكافأة تراكمية محددة (يتم تعينها عن طريق التجربة والخطأ) ويعطي عندها المؤشر (Is_Done) ثم يعتبر الوكيل الذكي مدربا إذا استقر المتوسط الحسابي لآخر 350 حلقة (Avg_Last350) على هذه القيمة الحدية وعندها يمكن اعتبار الوكيل قد تعلم وحل المهمة الموكلة إليه ويعطي عندها العلامة (Solved).
2) ثانيا، (الاختبار الثاني)، تم تكرار الاختبار الأول مع زيادة عدد الخطوات بالحلقة الواحدة إلى قيمة 2000 خطوة مما أدى إلى إجمالي 10 مليون خطوة تدريب.
3) لاحقا لذلك (الاختبار الثالث)، عُدّل هيكل المكافأة السابقة بالانتقال إلى الهيكل **B** مع الإبقاء على الإجراء **A**.
4) ثم (الاختبار الرابع) تم تدريب الوكيل من البداية مجددا باستخدام الإجراء **B** مع هيكل المكافأة **B**. مع تقليل عدد خطوات التدريب إلى 700 خطوة ضمن 5000 حلقة. اعتبرت الحلقة ناجحة إذا استطاعت تجاوز مكافأة تراكمية قدرها 180، وتدعىEp_Scr كما ذكر في الأعلى وهي المكافأة التراكمية للحلقة الواحدة، وقيمتها

الأعظمية = عدد الخطوات الكلي في الحلقة، لأن كل خطوة تنتج مكافأة تحسب بحسب تابع المكافأة وثم تقيس ضمن المجال [1-0] لذا، بفرض كل خطوة حققت قيمة 1 الأعظمية علما أنا هذا من الصعب جدا الوصول إليه لأنه يحتاج تفعيل كل الحساسات معا، فتكون القيمة الأعظمية لمجموع المكافآت في الحلقة Ep_Scr تساوي عدد الخطوات في الحلقة. في هذه الحالة يتم تعين قيمتها تجريبا من خلال مراقبة القيم الكمية لتابع المكافأة والقيم النوعية لشكل القبضة المتشكلة ضمن المحاكاة، وتبين أن 180 قيمة ملائمة لتشكل قبضة فعّالة وتقريبا هي أكثر قيمة تمكن الوكيل من الوصول إليها أثناء التجريب.

نهاية هذه المرحلة حُصل على السياسة النهائية المدربة والتي أجري عليها جملة من الاختبارات والتعديلات للأمثلة.

5.5.1.4. تقييم صلابة السياسة الناتجة وتعميمها (اختبارات المرحلة الثانية):

هدفت هذه المرحلة إلى اختبار صلابة السياسة المدربة والناتجة عن المرحلة الأولى، أي دراسة مدى قدرتها على التعميم في ظروف غير مألوفة أو متغيرة، دون إعادة تدريب، فقط بالاستفادة من الخبرة التي حصلت عليها سياسة اتخاذ القرار من التدريب السابق. بدأ الاختبار باستخدام نفس الإعدادات النهائية للبيئة والسياسة المدربة، ووفق معيار التقييم الأساسي (المكافأة التراكمية)، وبمتابعة منحني المتوسط المتحرك للمكافآت التراكمية عبر الحلقات التجريبية والتي بلغ عددها 100 حلقة تجريب (محاولة إمساك)). بعد ذلك نُفذت مجموعة الاختبارات التالية بالترتيب:

1) **استبدال الجسم:** تم استبدال علبة المشروبات المعدنية المستخدمة في التدريب بقارورة ماء، لاختبار قدرة السياسة على التعامل مع أشكال مماثلة لم يدرب عليها.
2) **زيادة التنوع في الأجسام:** تم اختبار الإمساك بقارورة زجاجية ذات شكل ووزن مختلفين.
3) **تعديل موضع الجسم أفقيا:** تم إزاحة العلبة الأصلية عدة سنتيمترات (أقل من 5 سم) على المحور الأفقي.
4) **إزالة الطاولة:** تم اختبار السياسة مع جعل الجسم متحركا عاموديا نحو الأسفل ومزاح أفقيا (السقوط العمودي).
5) **اختبار شكل قبضة مختلف:** استخدم جسم كروي (برتقالة) لاختبار شكل قبضة مغاير. وهنا تم تعديل وضعية اليد والجسم، لدراسة حالات أكثر تنوع.
6) **جعل الجسم الكروي متحركا:** تم تكرار التجربة مع برتقالة تتحرك عاموديا.
7) **استبدال الجسم بفاكهة ذات خصائص أخرى:** تم الاختبار مع تفاحة تملك خصائص شكلية وكتلية مختلفة.
8) **إزاحة التفاحة أفقيا** بمقدار 5 سم، ثم أكثر من 5 سم لاختبار التحمل أمام تغير الموضع.

تحدد هذه التجارب مدى مرونة السياسة المدربة، وقدرتها على تعميم الخبرة التدريبية على أجسام وبيئات غير مألوفة، دون الحاجة لإعادة التدريب أو تعديل المحددات. تعرض نتائج هذه الاختبارات في قسم النتائج.

5.5.1.5. تنفيذ تعلم المناهج (اختبارات المرحلة الثالثة):

في هذه المرحلة استُخدمت سياسة المرحلة السابقة كأساس لبناء تعلم متدرج عبر ما يعرف بـ "تعلم المناهج"، حيث تم متابعة تدريب النموذج على مهمة أكثر تعقيدا. الهدف من هذا الأسلوب هو توسيع قدرات النموذج تدريجيا بدءا من حالات أبسط (التي سبق أن درب عليها) وصولا إلى مهام أكثر تحديا. ثم تم اختبار السياسة الناتجة بعد تعلم المناهج على جسم مستطيل الشكل (حاوية بلاستيكية) ضمن Webots، يمثل ذلك تحولا جذريا عن شكل القبضة الأسطوانية التي تدرب عليها ما يعكس إذا حقق نتائج جيدة قدرة السياسة على التكيف مع اختلافات كبيرة.

5.5.1.6. اختبار قوة النظام المقترح على إنشاء قبضات مخصصة مستقبلا (المرحلة الرابعة):

الأساس في نماذج التعلم المعزز العميق هو بناء منهجية وهيكل مكافأة قوي يمكن لاحقا عند نشره على طرف حقيقي تدريبه على أجسام خاصة بنمط حياة المبتور بسهولة، تهدف هذه المرحلة إلى اختبار قوة المنهجية المقدمة وهيكل المكافأة المعتمد في التدريب، من خلال تطبيقه على مهام جديدة تماما بدءا من الصفر (أي بدون استخدام سياسة مدربة مسبقا). تم اختيار مهمة إمساك بمطرقة من بيئة Webots لتمثيل حالة واقعية من الاستخدام المستقبلي للطرف الاصطناعي. أجرت اختبارات المرحلة كما يلي:

1) الاختبار الأول: تدريب الوكيل من الصفر على إمساك مطرقة (قبضة أسطوانية أضيق) باستخدام: 1000 حلقة تدريب و700 خطوة لكل حلقة، مع حد نجاح للحلقة عند مكافأة تراكمية > 180.
2) الاختبار الثاني: لزيادة فعالية القبضة، تم رفع عدد الخطوات إلى 1500 وعيّن الحد الأدنى للمكافأة التراكمية إلى 350، لمدة 500 حلقة نظرا لأن شكل المطرقة يتطلب قبضة أدق وأكثر تعقيدا.
3) الاختبار الثالث: تم زيادة حد النجاح المطلوب إلى 600 لزيادة دقة الإمساك وأمثلة الجودة.
4) الاختبار الرابع: اختبار السياسة مع إزاحة المطرقة أفقيا عدة سنتيمترات لكشف قدرتها على التكيف المكاني بعد التدريب، إذ يتوقع عادة عند الاستخدام أن يزيح المستخدم الطرف عن الحالة الافتراضية للتدريب عدة سانتي مترات أفقيا أو عاموديا، وليس أكثر من ذلك.

5.5.2. حالة حساسات القوة:

التطبيقات العملية للأطراف الاصطناعية في سيناريوهات الإمساك ستتضمن تفاعلا فيزيائيا مباشرا مع البيئة، عند التعامل مع أجسام لينة وقابلة للتشوه بما فيه الكفاية يمكن أن يسبب الطرف تلف للجسم الملتقط وأذية، وكذلك الأمر في بعض الحالات قد يطبق الطرف قوة زائدة على جسم قوي وقاس أكثر من المتوقع فيتكسر الطرف عند تطبيق هذه القوى غير المدروسة، وفي حالات أخرى قد يفشل الوكيل في أداء المهمة بشكل كامل كمحاولة التقاط جسم ثقيل بقوة منخفضة أو العكس، وهنا تصبح القدرة على إدراك قوة التلامس والتحكم بها أمرا بالغ الأهمية. من هذا المنطلق، تهدف هذه الحالة الدراسية إلى استكشاف إمكانية تدريب وكيل التعلم معزز على تعديل سلوكه بناء على بيانات واردة من حساسات القوة، بما يتيح له اتخاذ قرارات تتسم بالحذر والدقة أثناء التفاعل مع بيئة حساسة.

5.5.2.1. صياغة المشكلة وفق عملية قرار ماركوف:

في هذه الحالة من البحث ومع استبدال حساسات اللمس بحساسات قوة، بحيث يتاح للوكيل تنظيم القوة المطبقة بشكل ديناميكي وملائم للسياق مما يزيد من واقعية النظام وكفاءته ستختلف مكونات MDP على النحو التالي:

1. **الوكيل:** لا يتم إجراء أي تغيير في هذا العنصر، حيث ما يزال يمثل الكيان القادر على اتخاذ القرار وتعلم سياسة مثلى باستخدام خوارزمية PPO.
2. **البيئة:** تم الاحتفاظ بنفس بيئة المحاكاة المبنية (يد روبوتية أمام علبة معدنية على طاولة باستخدام Webots، ولكن تم تعديل نوع الحساسات في الأصابع من حساسات اللمس الثنائي إلى حساسات القوة والتي تعيد قيم مستمرة تعبر عن مقدار الضغط (شدة القوة المطبقة من كل إصبع) بالنيوتن على الجسم المقابل. لأن البيئة هنا تتطلب ضبط القوة المطبقة لتحقيق الأمان والفعالية.
3. **الحالة المرصودة:** هي 20 حالة أيضا، ولكن هذه المرة ستصبح 20 قيمة مستمرة، على عكس الحالة السابقة والتي كانت 10 قيم مستمرة لزوايا المفاصل و10 قيم ثنائية لحساسات اللمس.

Observation Space (20) Continuous
10 Force Sensors 10 Joints Position

4. **الإجراءات:** سيتم التعامل مع الإجراء B فقط والموضح في حالة حساسات اللمس سابقا، والذي أثبت فعاليته في اختبارات الحالة السابقة.
5. **المكافآت:** هنا تم تعديل هيكل المكافأة بحيث يتم تضمين مقدار القوة المطبقة فيه وتابع المكافأة بهذه الحالة سمي C مكون من عدة أجزاء ويحسب المكافأة كالتالي:

- **أولا:** النسبة المئوية لعدد حساسات اللمس المفعلة ويتم اعتبار الحساس مفعل إذا تجاوزت القوة المقروءة منه قيمة 0.02 نيوتن لأن القيم أقل من هذا بالتجريب تبين أنها قد تكون ضجيج في الحساس وتعود منه كخرج حتى في حال عدم تصادمه مع شيء، علما أن هذا الضجيج يحسن مستقبلا من النشر.
- **ثانيا:** أُعطي الوكيل مكافأة صغيرة على رفع الجسم الممسوك عن الطاولة عدة ميلي مترات وكلما ثبت على هذا الارتفاع يقوم بتجميع مكافأة أكبر.
- **ثالثا:** تم خصم قيمة من المكافأة السابقة كعقوبة على الضغط الزائد. كلما ارتفع الضغط عن الحد المطلوب تزداد العقوبة بشكل طردي، أما إذا قل عن الحد المطلوب لرفع الجسم يأخذ عقوبة ثابتة.

Reward Function
RF = Percentage of enable Touch Sensors + Object Lift – Pressure Penalty Rescale [0 – 1]

6. **السياسة ودالة الانتقال:** لا تزال السياسة تعتمد على خوارزمية PPO لتطوير السياسة المثلى لهذا التفاعل أيضا بأسلوب الفاعل الناقد ولا يوجد تصريح لدالة الانتقال ويتم تعلمها مباشرة من خلال التفاعل مع البيئة.

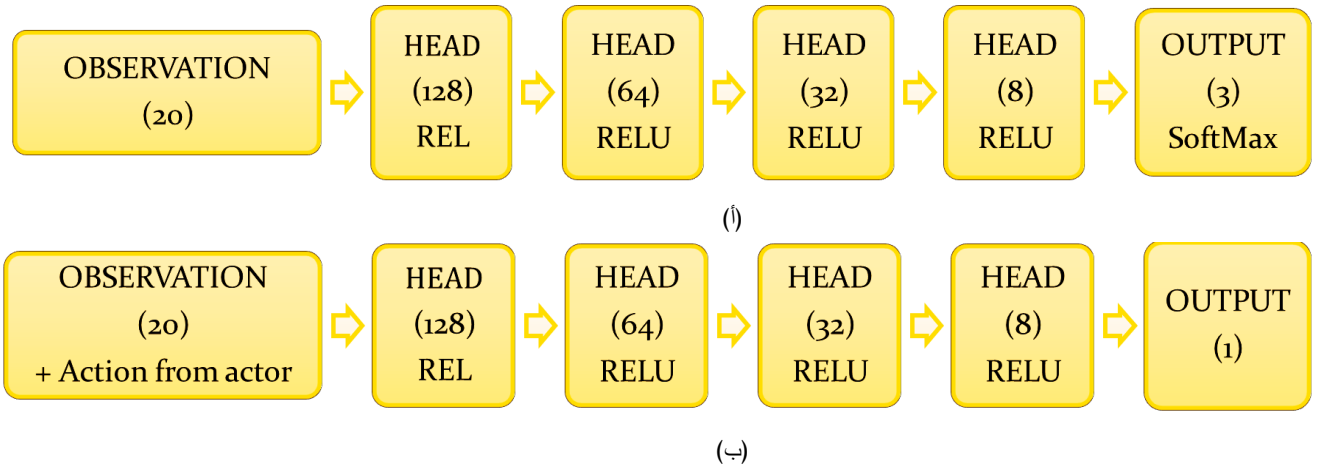
5.5.2.2. مقاييس الأداء لهذه الحالة:

يُدرس في هذه الحالة المقايضة بين هدفين، هما زيادة معدل النجاح وأمثلة قوة الإمساك. لذا لزم إضافة مقاييس إضافية للمقاييس السابقة إذ تضم هذه الحالة مقياس فرق لضغط (القوة الزائدة أو المنخفضة) وتم ترميزه بـ p ويحسب بالعلاقة، $p = F - W$ حيث F هي القوة المقاسة من المحاكاة والمطبقة على الجسم، وW تمثل وزن الجسم وهي تعبر عن أدنى قوة يجب أن تطبق لكي يتم إمساك الجسم وتحسب بجداء كتلة الجسم بتسارع الجاذبية الأرضية. تم حساب فرق الضغط هذا (القوة الزائدة أو المنخفضة) لكل خطوة زمنية ثم أُخذ المتوسط الحسابي لخطوات كل حلقة بالقيمة المطلقة وسمي فرق الضغط التراكمي P، وتمت مراقبته عن طريق المتوسط الحسابي المتحرك البسيط. يجب أن يعمل الوكيل على جعل هذا الفرق صفر أو قريب من الصفر.

5.5.2.3. بناء وتدريب سياسة اتخاذ القرار:

1. تعين بنية شبكة الفاعل – الناقد والمحددات الفائقة للنموذج المدرب:

في البداية استخدمت شبكتي الفاعل الناقد، والمحددات الفائقة كما هي في تجارب اللمس والمذكورة في الفقرة (5.5.1.3). ولكن مع تقدم التجارب لوحظ أن تعديل بعض المحددات وبنية شبكتي الفاعل – الناقد بعد عدة تجارب ساعدت في أمثلة النتائج. علما أن هذا التعديل احتاج مجموعة من التجارب اليدوية الكثيفة حتى تم اختيار القيم النهائية، وهي كالتالي: الشبكتان الجديدتان مصممتان كبنى بسيطة ولكن أعقد وأكثر عمقا، هذا العمق يحسن القدرة على التعلم خاصة في البيئات المعقدة. البنى الجديدة للشبكات تحتوي كل منهما على أربع طبقات مخفية تتكون الطبقة المخفية الأولى من 128 عقدة لكل طبقة (تحول الملاحظات الـ20 الخام إلى تمثيل أكثر تعقيدا ليسهل تعلم الأنماط المعقدة والمخفية ويسمى هذا بنهج التمثيل العالي المستوى، ثم تليها طبقة مع 64 عقدة، يليها 32 عقدة، ثم 8 عقد، مع استخدام تابع التفعيل ReLU بعد كل طبقة. شبكة الفاعل تنتهي بطبقة إخراج مع تابع Softmax لتوليد توزيع احتمالي للأفعال، بينما تنتهي شبكة الناقد بطبقة واحدة بدون تابع تفعيل لإخراج تقدير قيمة الحالة (متوقع المكافآت التراكمية من الحالة). تم استخدام محسن Adam لكل من شبكتي الفاعل والناقد، مع تحديد معدل تعلم قدره 0.001 للفاعل و0.003 للناقد وهي القيم الافتراضية التي يقدمها DEEPBOTS عند استدعاء PPO. أيضا يتم حفظ التحركات في ذاكرة انتقال مؤقتة (Replay Buffer)، ومن ثم تستخدم لتحديث النموذج بطريقة مصغّرة باستخدام مقاطع فرعية عشوائية (Mini-batch updates) من البيانات.



الشكل 21-5 بنية شبكتي الفاعل (أ) – الناقد (ب).

الجدول 5-5 المحددات الفائقة لخوارزمية PPO بعد التحديث.

المعامل	القيمة	الشرح
ε Clip-param	0.1	حد القطع للنسبة بين السياسات
max_grad_norm	0.3	أقصى قيمة لمقدار التدرج المسموح بها يمكن استخدامه لقص التدرجات أثناء الأمثلة لمنع "انفجار" التدرجات
PPO_update_iters	5	عدد تكرارات تحديث PPO في كل دورة
Batch -Size	10	حجم العينة المستخدمة في التحديث
γ Gamma	0.99	عامل الخصم لتقدير المكافأة التراكمية المستقبلية عادة يكون ضمن المجال [0-1] وبالتالي يحدد أهمية المكافأة المستقبلية. القيم القريبة من 1 تدل على إعطاء أولوية للمكافآت المستقبلية.

بالمقارنة مع الجدول (4-5) عملية تقليل معامل القطع تجعل التحديثات أكثر تحفظا، وبالتالي النظام أكثر استقرار لكن أبطأ في التعلم. كما أن تقليل معامل التدرج الأعظمي المسموح للتدرج يقيد النمو ويمنع من التحول الكبير "انفجار" السياسة ولكن أيضا يبطء التعلم، كما أن زيادة حجم الدفعة من المعروف أنه يعطي تحديث أكثر ثبات.

2. اختبارات التدريب وبناء السياسة (اختبارات المرحلة الأولى):

تم تدريب الوكيل ضمن البيئة على مهمة إمساك علبة المشروبات المعدنية، ولكن هذه المرة بأقل ضغط ممكن على العلبة بحيث نضمن عدم تشوه الجسم لاحقا، يتم قراءة قيمة الضغط الكلي المطبق على العلبة بجمع كل القيم العائدة من الحساسات العشرة في كل خطوة من الحلقة ويقارن مع قيمة 3.4 نيوتن إذ تعد هذه القيمة أدنى قيمة مطلوبة لإمساك العلبة المعدنية والتي تزن ضمن التطبيق 340 غرام فيكون حسب نيوتن ($F = M \times g = 0.33 \times 9.81 =$ 3.3354 N) ثم يتم حساب الفرق بين مقدار الضغط الكلي وأقل ضغط مطلوب وفق العلاقة ($P = Delta_Force = Total_Force - 3.4$) ويجب على الوكيل أن يتدرب بحيث يقلل هذه القيمة لأقصى قدر ممكن، ويقصد بأقل قوة للإمساك أقل قوة عمودية يجب تطبيقها بإصبعين على الجانبين أو أكثر ليمُنع الانزلاق الناتج عن الوزن للأسفل.

1) أولا، (الاختبار الأول) مع المحددات والشبكات السابقة ذاتها و5000 حلقة (تم التدريب عبر 600 منها فقط إذ أوقفت عملية التدريب يدويا) لكل منها 1500 خطوة، كما عين حد نجاح الحلقة الواحدة IS_DONE من خلال تحقيقها مكافأة تراكمية أكبر من 500، وتعتبر المهمة مكتملة والوكيل مدرب Solved إذا كان

المتوسط الحسابي لآخر 350 حلقة Avg_Last350 أكبر من 500. وهذه المرة بالإضافة لتتبع المتوسط الحسابي المتحرك للمكافآت التراكمية عبر الحلقات بنافذة 100 للمراقبة بحيث يجب أن يزداد مما يدل على التعلم، تم مراقبة المتوسط المتحرك لقيم فرق الضغط التراكمي المطبق والمأخوذ كمتوسط حسابي عبر الخطوات بالقيمة المطلقة كما ذكر في الأعلى، وهي القيمة التي يجب أن تتناقص أثناء التدريب على عكس المكافأة التراكمية التي يجب أن تزداد.

2) ثانيا، (الاختبار الثاني)، في تجربة لاحقة تمت متابعة تدريب السياسة الناتجة عن التجربة السابقة بمقدار 400 حلقة إضافية تقريبا.

تم تعديل قيم المحددات الفائقة عن طريق التجربة والخطأ وبنية شبكتي الفاعل الناقد كما عرضو في الأعلى ثم:

3) ثالثا، (الاختبار الثالث)، أعيد التدريب من الصفر مع 5000 حلقة و1500 خطوة لكل منها، مع حد نجاح 500 للحلقة وتعتبر المهمة مكتملة إذا كان المتوسط الحسابي لآخر 350 حلقة يساوي 500.
4) رابعا، (الاختبار الرابع)، تمت متابعة التدريب للنموذج السابق مع رفع حد الاستقرار من 500 إلى 800 وهي قيمة تجريبية لتعيين حد الاستقرار تُوقع أن زيادتها ستجعل الوكيل يتحسن في الأداء.
5) خامسا، ثم في (الاختبار الخامس)، وبهدف أمثلة النتائج تمت متابعة التدريب للنموذج السابق مع تعيين حد الاستقرار بالتجريب على 745.

5.5.2.4. تقييم صلابة السياسة الناتجة:

تهدف هذه المرحلة إلى اختبار صلابة السياسة المدربة والناتجة عن الاختبار السابق ودراسة مدى قدرتها على التعميم في ظروف غير مألوفة ومتغيرة دون إعادة تدريب فقط بالاستفادة من الخبرة التي حصلت عليها من التدريب السابق. بدأ الاختبار باستخدام نفس الإعدادات النهائية للبيئة والسياسة المدربة ووفق نفس معيار التقييم (المكافأة التراكمية) وبمتابعة منحني المتوسط المتحرك للمكافآت التراكمية عبر الحلقات التجريبية والتي بلغ عددها 100 حلقة تجريب (محاولة إمساك)) مع مراقبة فرق الضغط التراكمي P. بعد ذلك نُفيذ مجموعة اختبارات بالترتيب التالي:

1. **تعديل موضع الجسم الملتقط:** إزاحة علبة المشروبات المعدنية 5 سم أفقيا.
2. **تعديل موضع الجسم أفقيا وجعله متحرك:** تم إزاحة العلبة الأصلية 5 سم على المحور الأفقي، و**مع إزالة الطاولة**، تم اختبار السياسة مع جعل الجسم متحركا عاموديا نحو الأسفل (السقوط العمودي).
3. مع نفس الإعدادت للتجربة 2 ولكن رفع 10 سم تقريبا نحو الأعلى.
4. **تجربة أجسام مختلفة بذات الوزن مع انزياحات** أصغر من 5سم (الفازة 2 سم) الزجاجة 5 سم وأقل.
5. **شكل قبضة مختلف وتوضع الكف مختلف:** برتقال، بوزن 300 غرام.

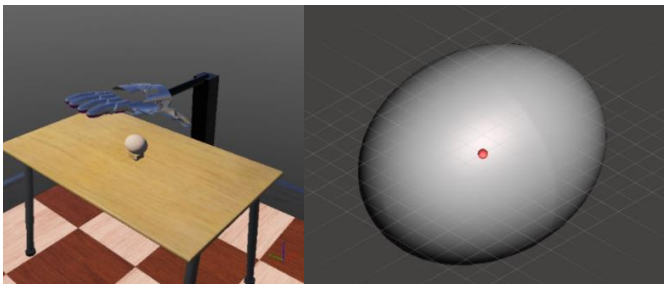
تحدد هذه التجارب مدى مرونة السياسة المدربة وقدرتها على تعميم خبرتاها على حالات جديدة دون الحاجة لإعادة التدريب أو تعديل المحددات. تعرض نتائج هذه الاختبارات في قسم النتائج ونتيجة لها تم الانتقال للتجارب اللاحقة.

5.5.2.5. تنفيذ تعلم المناهج:

في هذه المرحلة تمت متابعة تدريب النموذج الأخير في المرحلة السابقة على مهمة البرتقالة بهدف تطبيق تعلم المناهج لتعليم السياسة هذه المهمة بشكل متدرج، في البداية، درب على 5000 حلقة كل منها مع 1500 خطوة وحد نجاح 500، ثم تم إعادة التدريب ولكن بتقسيم عملية التدريب لمرحلتين، المرحلة الأولى درب ما يقارب 1800 حلقة، ثم في الجزء الثاني تمت متابعة التدريب ولكن الهدف كان تحسين الاستقرار من خلال زيادة الاستغلال وقمع الاستكشاف، لذا تمت تقليل قيمة معامل القطع clip_param=0.001، قبل متابعة التدريب مع ما يقارب 1000 حلقة إضافية حد نجاح الواحدة منها 500، وتعتبر المهمة منجزة إذا كان المتوسط الحسابي لآخر 100 حلقة يساوي 500، تقليل معامل القطع سيمنع المزيد من الاهتزاز بالتالي الاستقرار أسرع، وبما أن الاستكشاف كان كثيفا في الـ1800 حلقة السابقة فنريد هنا للنموذج فقط أن يستقر.

5.5.2.6. اختبار قوة النظام المقترح على إنشاء قبضات مخصصة مستقبلا (المرحلة الرابعة):

كما في حساسات اللمس يجب أن تختبر قوة المنهجية المقدمة، ولهذا الهدف دُرّب نموذج جديد من الصفر على مهمة جديدة كليا تتمثل في إمساك بيضة الدجاج بثلاث أصابع. ولإتمام هذه المهمة تم أولا تصميم بيضة الدجاج في برنامج Solidworks وفقا للمواصفات الموضحة في الجدول (6-5) وهي موضحة في الشكل (22-5)، ثم حفظت بصيغة STL وتم استيرادها لبيئة العمل حيث تم ضبط الكثافة للبيضة بالاستنتاج من الكثافة النوعية على القيمة $\rho = 1082\ kg/m^3$، ضمن بيئة Webots وزعت هذه الكثافة بالتساوي وبشكل متجانس وتمت مراقبة فرق الضغط المطبق عند الإمساك حيث تتأثر قوة كسر البيضة بوزنها وتتحمل بيضة الدجاج بوزن 64.07 غرام قوة 42 نيوتن للكسر (Altuntas et al, 2008) (Zhang et al, 2015).



الشكل 22-5 بيضة الدجاج المصممة في برنامج solidworks واستيرادها لبرنامج Webots.

الجدول 6-5 خصائص بيضة الدجاج المصممة (Joubrane et al, 2019).

الوزن g	الكثافة النوعية	سماكة القشرة مم	طول البيضة سم	عرض البيضة سم	ارتفاع الصفار مم	قطر الصفار سم	ارتفاع الألبومين مم
61.83	1.082	2.85	5.64	4.34	15.23	2.85	4.91

في المهمة يطلب من الطرف هنا محاولة إمساك بيضة الدجاج بثلاث أصابع فقط ومع 2000 حلقة لكل منها 2000 خطوة زمنية عيّن حد النجاح على القيمة 200 والمحددات كما ذكرت في الشكل (22-5) والجدول (5-6)، ثم تم الاختبار مع إزاحة 2 سم أفقيا، ثم تمت متابعة التدريب لمدة 1000 حلقة إضافية سعيا لتحسن الاستقرار.

5.6. الخاتمة:

أخيرا، عرض هذا الفصل خطوات التنفيذ العملي للبحث المقدم بشكل تفصيلي بدأ من تحضير الأدوات اللازمة للتطبيق العملي، ثم بناء وتجهيز بيئة التدريب وتحديد حالات الدراسة وشكل القبضات التي سيتم التدريب عليها، إذ تم التدريب على قبضة مستطيلة، وأخرى كروية، وأخرى أكثر دقة مع ثلاث أصابع، وصولا لتنفيذ كافة التجارب الخاصة بالبناء لسياسة اتخاذ القرار الأمثل واختبارها وأمثلتها من خلال تطبيق أسلوب تعلم المناهج، في الفصل اللاحق سيتم عرض نتائج هذه المنهجية بشكل مفصل، ومناقشتها ثم مقارنتها فيما بينها ومع الدراسات السابقة.

6. الفصل السادس: النتائج ومناقشتها.

6.1. تمهيد:

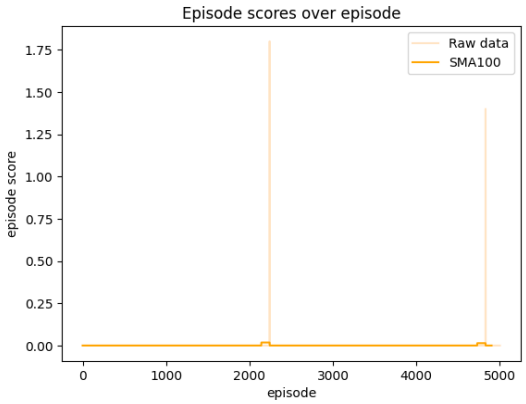
يصف هذا الفصل أولا النتائج الكمية والنوعية التي تم الحصول عليها لكل مكون من مكونات المراحل والتجارب السابقة، ومناقشتها لكل حالة على حدى، ثم يقارن النتائج فيما بينها لتبيان الفرق بين حساسات اللمس وحساسات القوة، ومن ثم مقارنة النتائج مع الدراسات السابقة ذات الصلة، وأخيرا يُعرض ملخص لهذه النتائج.

6.2. نتائج حالة حساسات اللمس (الحالة الأولى):

6.2.1. نتائج اختبارات بناء السياسة والتدريب (المرحلة الأولى):

أثناء التجارب في هذه المرحلة تم استكشاف فعالية هياكل المكافأة المختلفة والإجراءات المرتبطة بها من خلال التدريب على مهمة التقاط جسم أسطواني (علبة مشروبات معدنية).

1) في الاختبار الأول وبعد تدريب دام أكثر من 30 ساعة لم يستطع الوكيل الذكي الوصول لأي إجراء يعظم قيمة المكافأة التراكمية كما هو موضح في الشكل (1-6) كل الحلقات تأخذ قيمة 0 كمكافأة تراكمية، ونتيجة ذلك ربما يكون عدد الخطوات غير كافٍ لذا، تم تكرار التجربة السابقة مع زياد عدد الخطوات.
2) في نتائج الاختبار الثاني استطاع الوكيل تحقيق مكافأة تراكمية أكبر من الإعدادات السابقة كما يوضح الشكل (2-6) حيث بلغت تقريبا القيمة 15، ورغم تزايدها إلا أنها لم تكن كافية للوصول للحد المطلوب الذي يشكل قبضة فعالة وأخذ التدريب أكثر من 80 ساعة.
3) مع الاختبار الثالث وجد تحسن ملحوظ بقيمة المكافأة التراكمية وصلت حتى 250 تقريبا كما يوضح الشكل (3-6) وكانت تبدي في المقاييس النوعية قبضات ملائمة لحمل علبة المشروبات المعدنية إلا أن التدريب استغرق أكثر من 50 ساعة. ولتقليل وقت التدريب أجري الاختبار الرابع. (عينت EP_SCR على 1800 فقط من أجل تعيين أعظم قيمة يمكن للوكيل الوصول إليها).
4) في الاختبار الرابع استطاع الوكيل المدرب الوصول للحالة المستقرة بعد أقل من ساعة، ورغم ذلك درب ما يقارب الساعتين للاستقرار، كما يوضح الشكل (4-6). يعرض الجدول (1-6) ملخصا لهذه الاختبارات.



الشكل 1-6 مخرجات الإعدادات الأولية البيانات الخام RAW DATA لقيم كل epoch في حين تمثل SMA المتوسط المتحرك البسيط بنافذة 100.



الشكل 6–2 المخرجات عند زيادة عدد الخطوات.



الشكل 6–3 يوضح تحقيق الوكيل مكافأة أكبر عند استبدال تابع المكافأة.

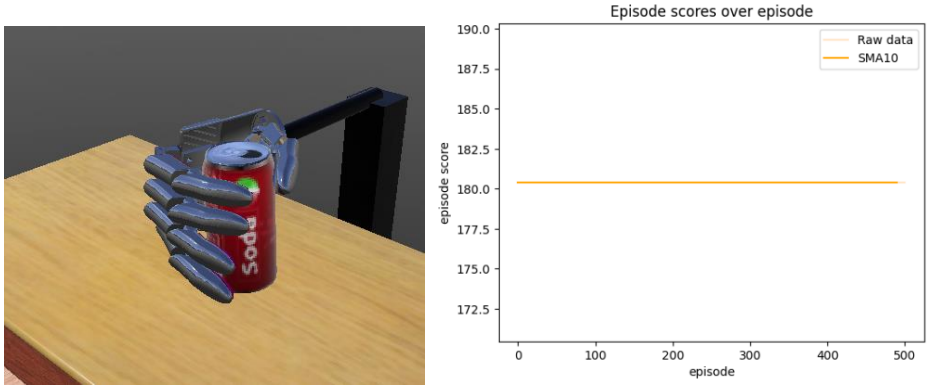


الشكل 6–4 وصول النموذج بالإعدادات النهائية للحالة المرغوبة.

الجدول 6–1 ملخص نتائج تجارب البناء والتدريب للسياسة في حالة حساسات اللمس.

الاختبار	الإجراء المستخدم	هيكل المكافأة	عدد الحلقات الكلي	عدد الخطوات لكل حلقة	إجمالي عدد الخطوات (مليون)	شرط نجاح الحلقة IS_DONE	شرط إكمال المهمة SOLVED	الزمن الكلي المستغرق	أعلى مكافأة تراكمية تم تحقيقها	ملاحظات
1	A	A	5000	700	3.5	EP_SCR >180	Avg_Last 350> 180	أكثر من 30 ساعة	0	لا يوجد تعلم فعال، المكافأة 0.
2	A	A	5000	2000	10	EP_SCR >180	Avg_Last 350> 180	أكثر من 88 ساعة	≈ 15	تحسن طفيف لكن غير كاف.
3	A	B	5000	2000	10	EP_SCR >1800	Avg_Last 350> 1800	أكثر من 50 ساعة	≈ 250	أداء جيد قبضات مناسبة.
4	B	B	5000	700	3.5	EP_SCR >180	Avg_Last 350> 180	< 1 ساعة	> 180 (قيمة الاستقرار)	نتائج جيدة بوقت قصير.

وعليه إن أفضل النتائج تم الحصول عليها عند استخدام هيكل المكافأة B مع الإجراء B، ضمن إعدادات تدريب مكونة من 5000 حلقة تدريب وكل حلقة مؤلفة من 700 خطوة، أدى إلى تحقيق مكافأة تراكمية تدريجية استقرت عند قيمة 180. اعتبرت هذه القيمة مؤشرا على تحقيق القبضة المثلى (حيث عينت القيمة بمبدأ التجربة والخطأ). كما بين الشكل (4-6) السابق أن الوكيل تمكن من الوصول إلى الأداء المرغوب بأقل من 400 حلقة، ما يعكس فعالية هذا التكوين الذي يجمع هيكل المكافأة B مع الإجراء B في تسريع التعلم. ومن المنطق أن يكون هذا الإعداد أفضل من الاقتراحات الأخرى لأنه يقلل عدد الإجراءات التي يجب على السياسة التعامل معها ويلغي الإجراءات التي لم تبدي أي فائدة والمتمثلة بإرجاع الزاوية بمقدار C كما أن هيكل المكافأة B والذي يعطي مكافآت متناسبة طردا مع عدد الحساسات المفعلة شجع السياسة على عملية الإمساك على عكس هيكل المكافأة A والذي يعطي مكافأة ثابتة فقط على زيادة تفعيل الحساسات فيكون دفع السياسة للتعلم أبطأ بكثير وقد يكون شبه معدوم كما عرضت النتائج. ثم اختبرت السياسة النهائية الناتجة على ذات الإعدادات وحققت على مدى 500 تجربة إمساك معدل نجاح 100%، وكما توضح النتائج العينية والكمية في الشكل (5-6)، حققت هذه السياسة في الاختبار أداءا مستقر بشكل واضح ما يدل على تدربها فعلا. بناء على ذلك، تم اعتماد السياسة الناتجة عن التجربة الأخيرة هذه كسياسة مدربة مُثلى نهائية، نتيجة توازن الأداء الجيد وفعالية وقت التدريب. ثم في المراحل التالية، أجريت تجارب التعميم والصلابة لاختبار مدى قوة السياسة المدربة أمام ظروف غير مألوفة.

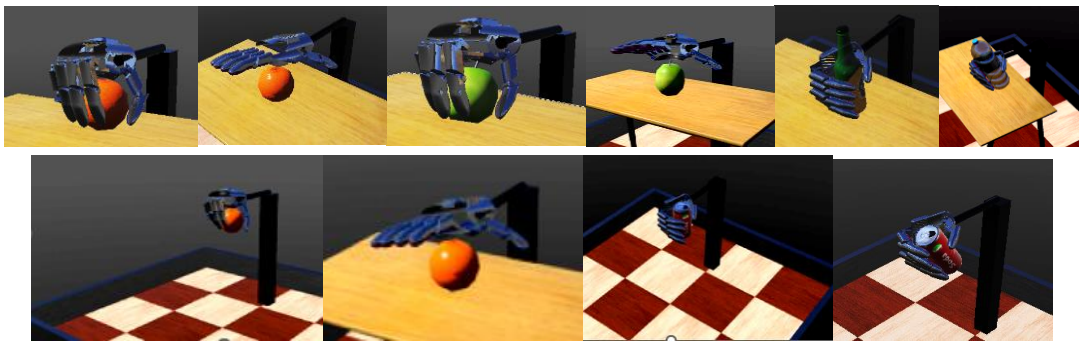


الشكل 5-6 النتائج الكمية والنوعية لاختبار السياسة الناتجة عن تجارب البناء والتدريب لحالة حساسات اللمس.

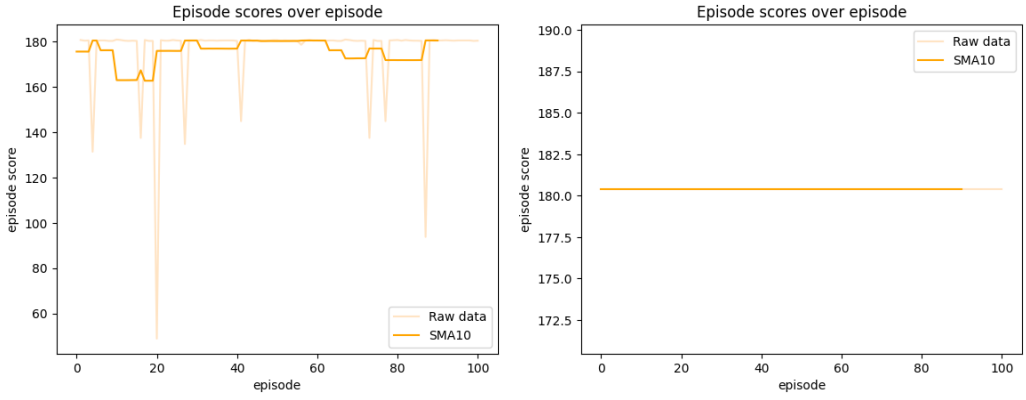
6.2.2. نتائج اختبارات التعميم والصلابة (المرحلة الثانية):

في هذه المرحلة كما تم التوضيح سابقا، بعد تثبيت السياسة الناتجة من مرحلة البناء والتدريب، تم إخضاعها لاختبارات صلابة شملت حالات متعددة لتقييم مدى قدرتها على التعميم على أجسام وبيئات جديدة. أظهرت النتائج أن السياسة حققت على مدى 100 تجربة، نسبة نجاح إمساك 100% في الإمساك بالأجسام المشابهة (قبضة أسطوانية) والمختلفة قليلا من حيث الحجم مثل قارورة ماء بلاستيكية وزجاجية. وأثار الاهتمام قيام السياسة بإمالة القارورة على المحاكاة كما يوضح الشكل (5-6) والشكل (6-6) أيضا مما يقلل خطر انزلاقها وكذلك قيمة جهد الإمساك. عند تعديل موضع الجسم أفقيا بالإزاحة أظهرت السياسة قدرة واضحة على التكيف مع التغيرات الموضعية

للجسم وحققت أيضا نسبة نجاح في الإمساك 100% وتعتبر هذه الحالة مهمة لأن المبتور المرتدي للطرف ربما يقوم بالإنزياح قليلا عن الحالة الافتراضية أو البعد المثالي الذي درب عليه الوكيل، كما حققت السياسة نجاحا بنسبة 91% عند إزاحة الجسم أفقيا وجعله عائم (إزالة الطاولة) ويدل ذلك على سرعة قبضة الطرف فهنا يقوم الطرف بتطبيق عزم فوري قوي بخطوات قليلة للحصول على القبضة المرغوبة بأجزاء من الثانية ويحافظ عليها باقي خطوات الحلقة، وعند إعادة ذات التجربة بتغير ارتفاع العبوة المعدنية الأولي بالرفع عدة سنتي مترات أعاد تحقيق نسبة 100% إذ تنخفض النسبة الأولى لأنه في بعض التجارب لايتمكن من إمساك العبوة قبل سقوطها وساهم الرفع بتحسين النتائج والتمكن من الإمساك قبل سقوط العبوة ويوضح الشكل (6–7) الفرق بين التجربتين عند تحقيق 100% وحالة 91%. عند اختبار شكل قبضة مختلف (القبضة الكروية) كان من اللافت أن السياسة تمكنت من التكيف مع أجسام مختلفة في الشكل، كبرتقالة وتفاحة، حتى مع وجود حركة عمودية أو إزاحة أفقية بسيطة بنسب نجاح 100% وكانت النتائج مستقرة أيضا في هذه التجارب. ومع ذلك، فشلت السياسة في التقاط جسم كروي (تفاحة) مزاح أفقيا أكثر من 5 سم، حيث حققت نسبة نجاح 0% وكان السبب أن الجسم الكروي الملتقط يقع بموضع بعيد نوعا ما عن الأصابع الأربع المتوازية وأقرب للإبهام، مما أدى لفشل عملية الإمساك، وكشف عن حدود التعميم في النموذج الحالي، وهو ما تمت معالجته لاحقا عبر تعلم المناهج. رغم أن المبتور منطقيا لن يقوم بمحاولة التقاط الجسم وهو بعيد عنه بشكل كبير إلا أن الهدف من هذه التجربة تضمين هذه الخبرة في التجارب اللاحقة. ويوضح الشكل (6–6) هذه النتائج العينية.



الشكل 6–6 النتائج العينية لاختبارات المرحلة الثانية في حالة حساسات اللمس.

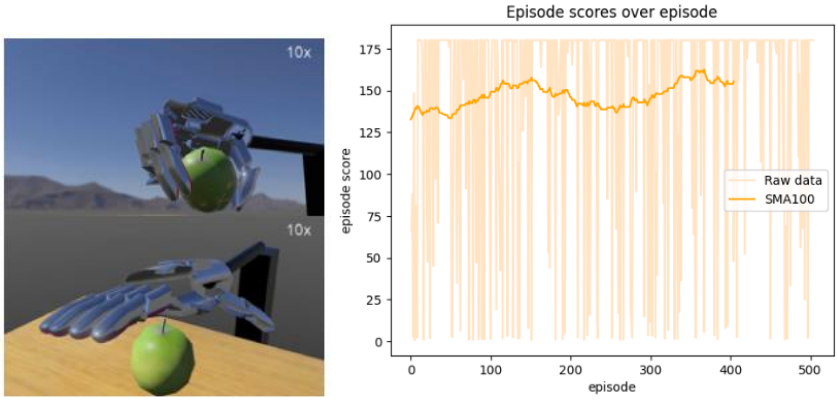


(ب) (أ)

الشكل 7–6 منحني المكافآت التراكمية لـ100 تجربة حققت 100% (أ) كنسبة نجاح في الإمساك، والأخرى (ب) حققت 91%.

6.2.3.نتائج تجارب اختبار تعلم المناهج (المرحلة الثالثة):

هنا استخدم أسلوب "تعلم المناهج" لتعزيز السياسة المدربة وإكسابها قدرة على التعامل مع التحديات الجديدة مثل الأجسام المزاحة بشكل كبير أو الأشكال الجديدة كليا. وذلك بهدف توسيع محدوديات النموذج التي ظهرت في المرحلة السابقة، أظهر النموذج تحسنا ملحوظا، إذ لوحظ نجاح الوكيل بعد 150 حلقة تدريب إضافية والتي انخفضت لتعود وتعطي نتائج أفضل بعد 500 حلقة تدريب، كما يوضح الشكل (8-6)، حيث تبين أن سياسة اتخاذ القرار لم تستقر بشكل كبير ولكن في الاختبار ارتفعت نسبة النجاح في التقاط الجسم الكروي المزاح من 0% إلى 100% وكانت مستقرة على مدى 100 تجربة بعد دمج الخبرات السابقة وتوسيع بيانات التدريب، لذا لم يتم أي تدريب إضافي بعد هذه المرحلة، ما يؤكد أن تعلم المناهج كان له دور كبير في أمثلة النتائج وتضمين خبرات جديدة بطريقة منهجية منظمة للسياسات. ولاحقا لذلك في محاولة لاختبار قدرة المتحكم الناتج تم اختبار السياسة الناتجة على التعامل مع مشاكل جديدة كليا حيث اختبرت على جسم مستطيل الشكل (حاوية فاكهة معروفة ضمن بيئة Webots) موضح في الشكل (9-6) ونجحت السياسة في التقاطه بنسبة 61%، مما يدل على تحسن قدرة النموذج على التعميم ومعالجة التحديات غير المسبوقة، وتبين حتى هنا قدرة السياسة الناتجة على التعامل مع الكثير من الإعدادات وإمكانية الاستفادة من تعلم المناهج، ولكن حتى هنا أوقفت هذه التجارب وتمت المتابعة للاختبار مع مهمة جديدة.




```
Episode # 498 score: 180.00000000000068
Episode # 499 score: 180.20000000000152
Episode # 500 score: 180.10000000000173
Episode # 501 score: 180.3000000000015
Episode # 502 score: 180.3000000000015
Task is solved, deploying agent for testing...
INFO: 'robot_supervisor_controer' controller exited successfully.
```

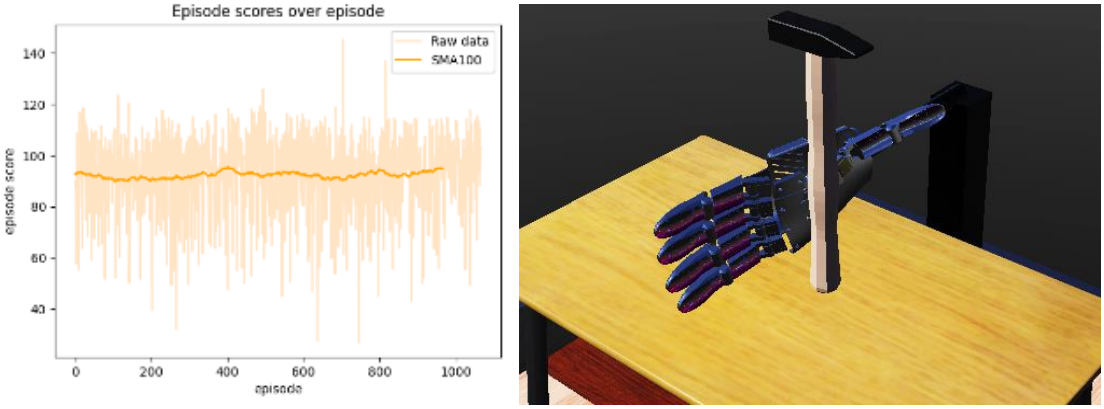
الشكل 8-6 نتائج تطبيق تعلم المناهج.



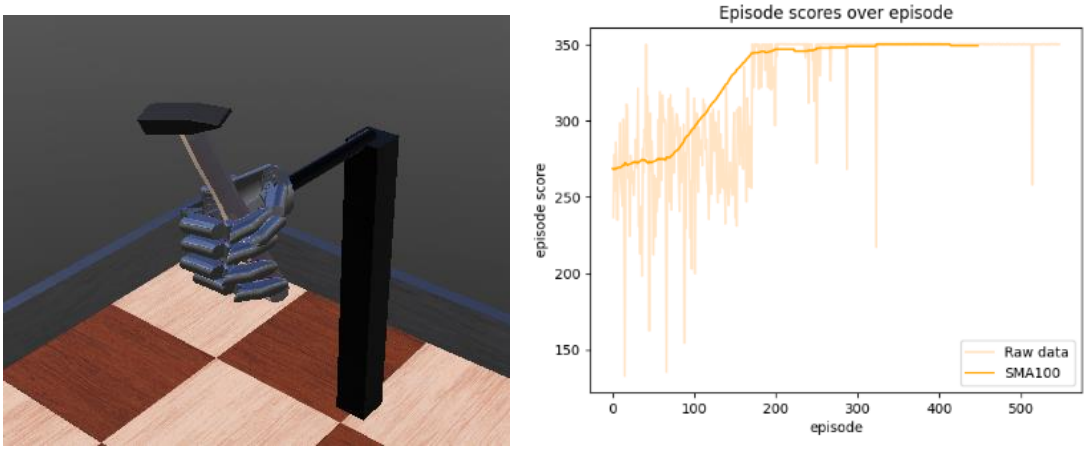
الشكل 9-6 نتائج الاختبار على جسم غريب كليا.

6.2.4. نتائج تجارب اختبار قوة النظام المقترح (المرحلة الرابعة):

ركز هذا القسم على تقييم بنية السياسة وقوة هيكل المكافأة المقترح عند استخدامه لتدريب نموذج جديد من الصفر على أجسام جديدة تماما. بهدف دراسة قوة النظام المقترح لأن الهدف النهائي الحصول على منهجية وهيكل مكافأة قوي بما يكفي لتدريبه على التعامل مع أجسام جديدة إذ يتم عادة في هذه النماذج تدريبها على أجسام خاصة بالمريض ونمط حياته. تم اختيار مطرقة كمثال على جسم له شكل قبضة مختلف (أضيق)، وبينت التجارب أن إعدادات التدريب السابقة لم تكن كافية لتحقيق الإمساك الناجح كما يوضح الشكل (10–6)، ويعتقد أن ضيق القبضة يحتاج عدد خطوات أكثر للوصول للحالة المثالية لذا تم زيادة عدد الخطوات لكل حلقة، وعند زيادة عدد الخطوات لكل حلقة إلى 1500 وتعيين حد المكافأة التراكمية إلى 350، نجحت السياسة في تحقيق قبضة فعالة كما يوضح الشكل (11–6). لاحقا لذلك تم اختبار صلابة النموذج المدرب في حالات متعددة: مع حركة عمودية للجسم، ثم مع زيادة حد المكافأة إلى 600 وتدريب 400 حلقة تقريبا لتحسين الاستقرار استطاع النموذج الاستقرار على 600، وأخيرا مع عدة اختبارات إزاحة. أظهرت السياسة في جميعها نسبة نجاح 100%، ما يعزز فرضية أن النظام المقترح قادر على دعم تعلم مرن وفعال لأنماط قبض جديدة عند التدريب من الصفر، لهذه التجربة أهمية كبيرة لاحقا تساهم في تدريب السياسة على أجسام خاصة بنوعية حياة المبتور وتؤكد النتائج أن المنهجية المقترحة يمكنها تحقيق ذلك، مع تعديلات بسيطة فقط في عدد الخطوات وحد النجاح والقليل من التدريب.



الشكل 10–6 نتائج التدريب على الإمساك بمطرقة بالإعدادات الأولية.



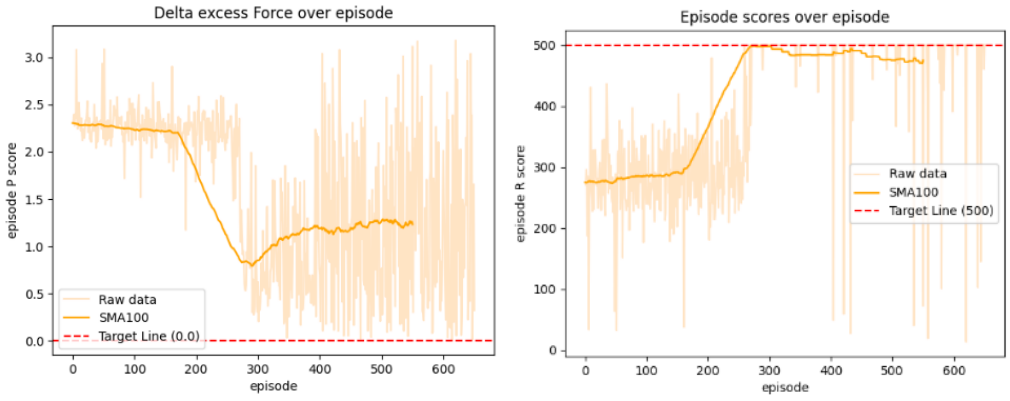
الشكل 11–6 نتائج التدريب الكمية والاختبار العينية على إمساك المطرقة مع زيادة عدد الخطوات.

6.3. نتائج حالة حساسات القوة (الحالة الثانية):

6.3.1. نتائج اختبارات مرحلة بناء السياسة وتدريبها (المرحلة الأولى):

أجريت في هذه المرحلة عدة اختبارات سيتم عرض نتائجها بالتسلسل وصولا إلى السياسة النهائية واختبارها:

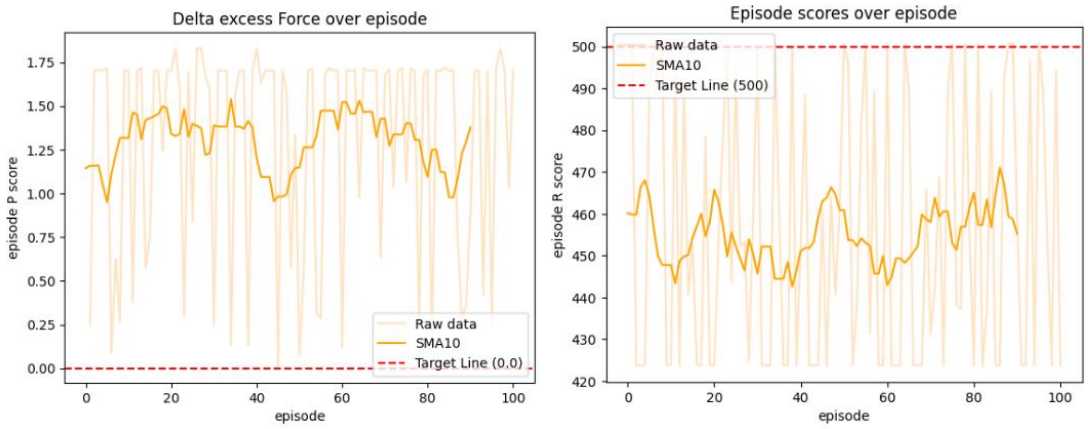
1) **نتائج الاختبار الأول:** أثناء التدريب أبدت هذه التجربة تزايد واضح بمنحني المكافأة وتناقص جيد بقيم فرق الضغط كما يوضح الشكل (6–12). عند اختبار السياسة الناتجة عن هذه المرحلة عبر 100 تجربة كما هو موضح في الشكل (6–13)، حقق الوكيل مع هذه السياسة نسبة نجاح في الإمساك Success rate test = 16% فقط، ووصل متوسط فرق الضغط لخطوات كل حلقة بالقيمة المطلقة وهو ما سمي فرق الضغط التراكمي P إلى قيمة أعظمية تساوي 1.75 N والذي يجب أن يكون صفر، –أخذت القيمة المطلقة لأن، الزيادة أو النقصان بالضغط مرفوض لذا تمت مراقبة الفرق والذي تسعى السياسة لتصفيره لا يهم إن كان زيادة أو نقصان– إذا منحني فرق الضغط الظاهر في منحنيات فرق الضغط للتدريب في الشكل (6–12 (ب)) أو الاختبار في الشكل (6–13 (ب))، هي قيم فرق الضغط التراكمي كمتوسط خلال 1500 خطوة بالقيمة المطلقة لكل حلقة، أما عند عرض قيمته لحلقة واحدة لا على التعيين كما في الشكل (6–14)، يلاحظ أن الضغط في البداية يبدأ سالب وهذا منطقي لأن عائد الحساسات في البداية قبل التصادم يكون معدوم وبطرح مجموع عوائد الحساسات من القوة الدنيا اللازمة سيكون الناتج سالب ومع تقدم الحلقات قد تصطدم بعض الأصابع بالعبوة فيتناقص الفرق إلا أنه لا يصل إلى 0 مما يدل على أنه في الحلقة المبينة في الشكل (6–14) مثلا سيفشل في رفع العبوة. كما لوحظ من منحني الضغط للخطوة أنه بمجرد بدأ التلامس تقريبا عند الخطوة 200 يقوم الوكيل المدرب مباشرة بتقليل الضغط لأن السياسة كما قلنا تقمع التغيرات الحادة، فيتراجع الضغط المطبق ويزداد الفرق مرة أخرى بشكل قليل ثم يستقر.



(ب) (أ)

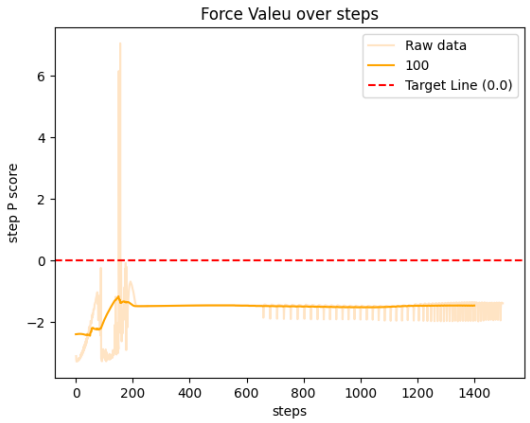
الشكل 6–12 منحنيات التدريب للاختبار الأول في حالة حساسات القوة

(أ) المتوسط الحسابي المتحرك للمكافآت التراكمية و(ب) المتوسط المتحرك لفرق الضغط التراكمي بالقيمة المطلقة عبر الحلقات.



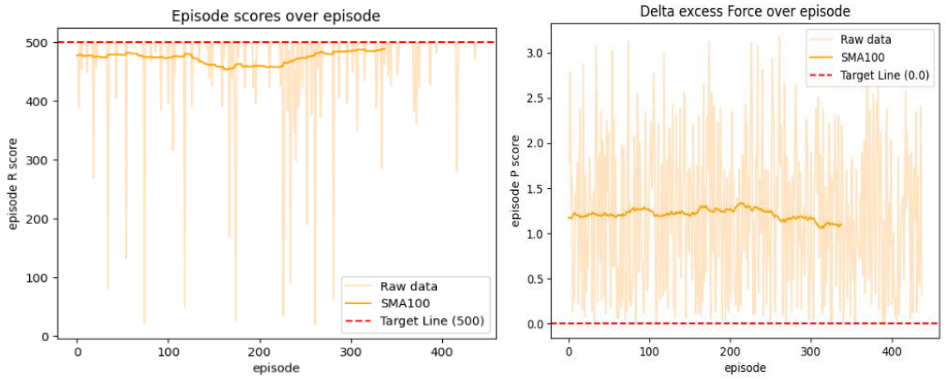
(ب) (أ)

الشكل 13–6 منحنيات تابع المكافأة التراكمية عبر 100 حلقة تجريب (أ)، و (ب) منحنيات فرق الضغط التراكمي بالقيمة المطلقة عبر 100 حلقة.



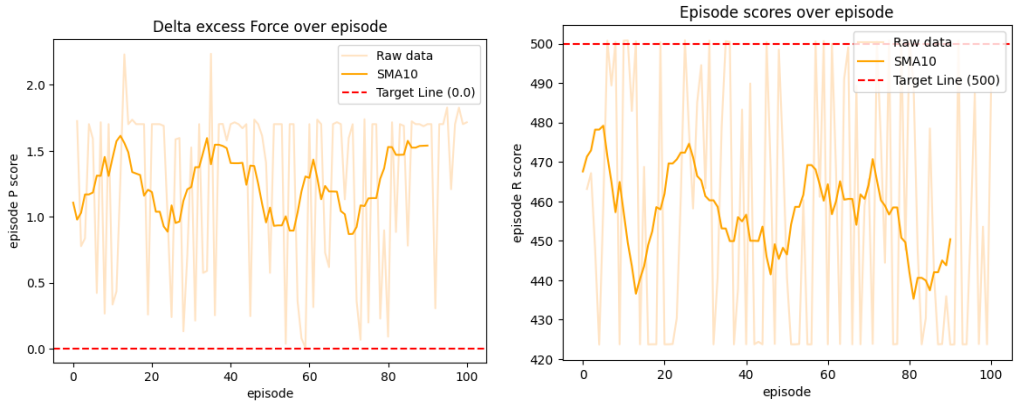
الشكل 14–6 قيمة القوة المطبقة في كل خطوة من أحد حلقات التجريب الفاشلة في عملية الإمساك.

2) **نتائج الاختبار الثاني:** عند التدريب لم يظهر تحسن كبير في النتائج ولم يستقر النموذج. فكما هو موضح في الشكل (15–6)، حيث يظهر النموذج الثاني المدرب أكثر نفس حدود الاستقرار تقريبا عند التدريب. وعند اختبار السياسة الناتجة عن هذا الاختبار على مدى 100 تجربة كما يوضح الشكل (16–6)، انخفضت قليلا قيمة P المطبقة حيث بلغ أعظمه 1.5 N وبلغ معدل النجاح = Success rate test 18%، أما في منحني الخطوة الواحدة الشكل (17–6 (أ))، يظهر التجريب لحلقة واحدة نتائج مقاربة جدا للتجربة السابقة في حالات الفشل المشابهة (عند تحقيق مكافأة منخفضة جدا)، عند إظهار المنحني لحلقة أخرى ناجحة (أي حققت المكافأة التراكمية 500)، الشكل (17–6 (ب))، لوحظ أنه عديم الاستقرار لكنه تمكن من الوصول لفرق ضغط صفر لفترة زمنية معينة ثم أعطى قيم ضغط أكبر بمقدار قليل لا يشكل فرقا شاسعا في التطبيق العملي إلا أنه انهار بعدها. في حلقة أخرى حققت مكافأة تراكمية قريبة من 500 لكنها لم تصل لحد النجاح لذا تعتبر فاشلة الشكل (17–6 (ج))، كما يمكن ملاحظة أنه حقق فرق الضغط قيمة 0 وأكثر قليلا من الصفر لفترة زمنية (عدد من الخطوات في الحلقة) أقل من الفترة السابقة. تدل هذه النتائج على تحسن طفيف جدا بالنتائج.



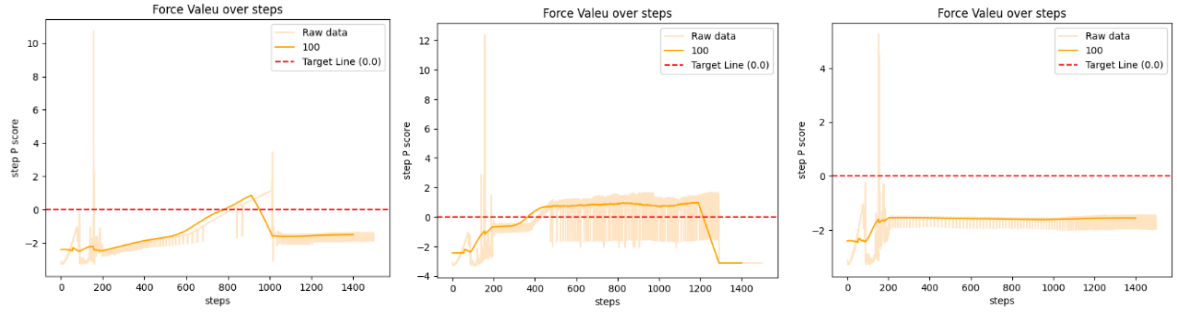
(ب) (أ)

الشكل 15–6 منحنيات التدريب للاختبار الثاني من المرحلة الأولى، (أ) منحني المكافآت، (ب) منحني فرق الضغط التراكمي



(ب) (أ)

الشكل 16–6 نتائج اختبار السياسة الناتجة عن الاختبار الثاني من مرحلة البناء والتدريب.



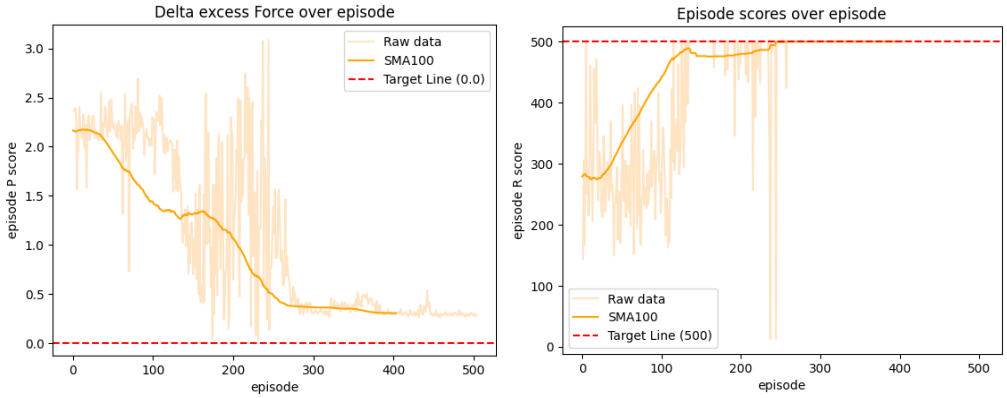
(ج) (ب) (أ)

الشكل 17–6 منحني فرق الضغط عبر الخطوات لثلاث حلقات من حلقات الاختبار للسياسة، (أ). حلقة فاشلة حققت أقل من 500 بكثير كمكافأة تراكمية. (ب). حلقة ناجحة حققت 500 كمكافأة تراكمية. و(ج). حلقة غير ناجحة لكنها حققت قيمة قريبة جدا من 500.

بشكل عام كانت السياسات الناتجة عن هذا الضبط عديمة الفائدة وتأخذ وقت كبير جدا ولا تحقق النتيجة المرغوبة لذا عُدلت المحددات الفائقة وفق ما ذكر في الأعلى، بزيادة عمق الشبكتين لأن المدخلات أصبحت أكثر تعقيدا فهي الآن 20 حالة مستمرة، كما تم زيادة مقدار العقوبة عندما يختلف فرق الضغط عن صفر.

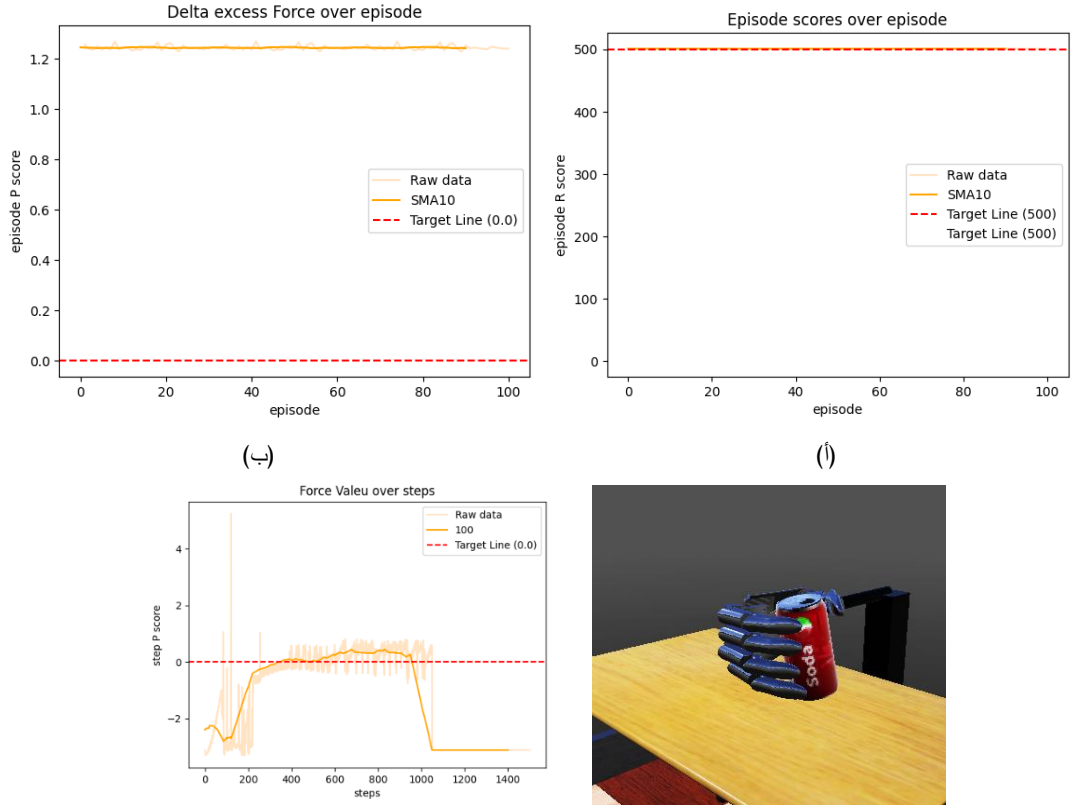
3) **نتائج الاختبار الثالث:** في التدريب لوحظ تزايد في قيمة المكافأة التراكمية واستقر النموذج المدرب على القيمة 500 بعد ما يقارب 500 حلقة تدريب وحقق التقارب بشكل أكبر كثيرا من الحالات السابقة، وأما منحني متوسط قيمة فرق الضغط التراكمي بالقيمة المطلقة استطاع أن ينخفض إلى قيمة أكبر قليلا من

0.5 نيوتن كما يوضح الشكل (18–6). عند اختبار السياسة الناتجة عن هذه الحالة، حققت 100 حلقة من التجريب مكافأة قيمتها 500 بشكل مستقر أي Succes rate test = 100%، كما يوضح الشكل (6–19 (أ))، أما فرق الضغط التراكمي كمتوسط بالقيمة المطلقة كان شبه مستقر على القيمة 1.2 نيوتن كما يوضح الشكل (19–6 (ب)). بالنسبة لنتائج الخطوة الواحدة كما يوضح الشكل (19–6 (د)) لوحظ تناقص ثم استقرار قيمة فرق الضغط لفترة طويلة على القيمة 0 إلا أنه كان غير مستقر حيث انهار في نهاية الحلقة وأعاد تطبيق فرق ضغط كبير ويكون السبب لهذه الحالة عدم كفاية التدريب لذا تم الاختبار الرابع.



(ب) (أ)

الشكل 18–6 منحنيات التدريب للاختبار الثالث من مرحلة التدريب، (أ) المتوسط المتحرك للمكافآت التراكمية، و(ب) متوسط متحرك لفرق الضغط التراكمي.



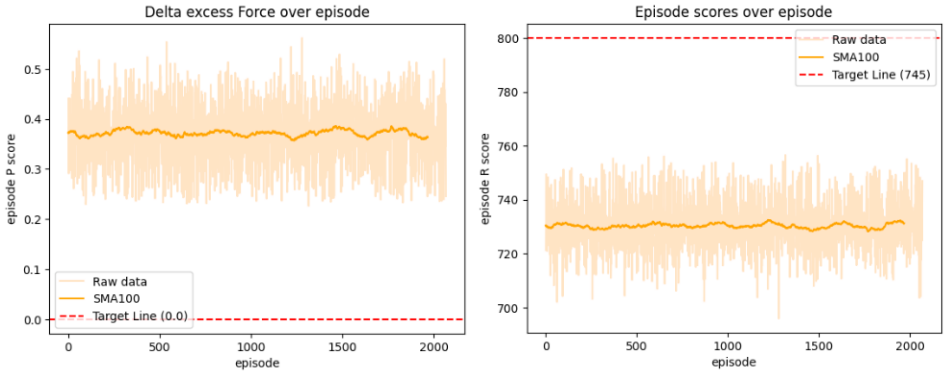
(ب) (أ)

(د) (ج)

الشكل 19–6 نتائج اختبار السياسة الناتجة عن الاختبار الثالث لمرحلة البناء

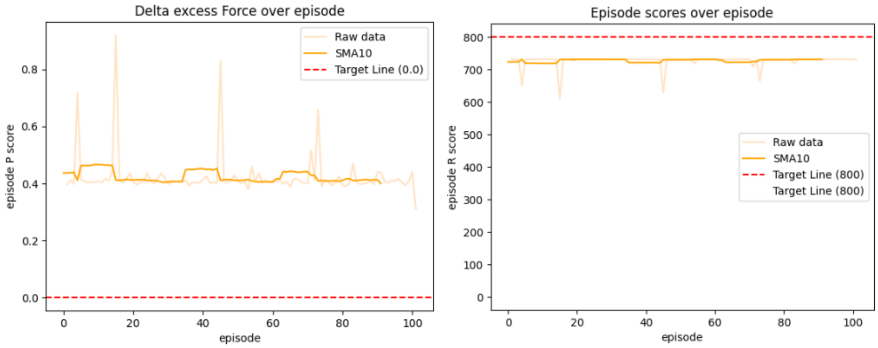
(أ). المكافأة التراكمية. (ب). فرق الضغط التراكمي، (ج). النتيجة العينية، (د) النتيجة لحلقة اختبار واحدة.

4) **نتائج الاختبار الرابع:** عند التدريب، يبين الشكل (6–20 (أ)) نتائج عملية التدريب للاختبار الرابع حيث تكون أقصى مكافأة تراكمية يمكن أن يحققها الوكيل تقارب 730 لذا اعتبر حد النجاح، بالنسبة لقيمة فرق الضغط التراكمي كما يوضح الشكل (6–20 (ب))، حقق استقر تقريبا على قيمة أقل من 0.4 نيوتن. وعند اختبار السياسة الناتجة عن الاختبار الرابع في هذه المرحلة، على 100 تجربة ومع حد نجاح 730 لأن 800 لم تصل له أي حلقة، كانت نسبة النجاح Success rate test = 92% وشبه مستقر، وأما قيمة متوسط فرق الضغط التراكمي 0.5 نيوتن كما يوضح الشكل (6–21)، إلا أنه بالنسبة للحلقة الواحدة الناجحة كان تقريبا مستقر على قيمة 0 أو أكثر قليلا، وأيضا في نهاية الحلقة عند آخر 28 خطوة تقريبا يقوم بإفلات العبوة كما يوضح الشكل (6–22).



(ب) (أ)

الشكل 6–20 منحنيات التدريب للاختبار الرابع من المرحلة الأولى. (أ).منحنيات المكافأة التراكمية، (ب) منحنيات فرق الضغط التراكمي كمتوسط.



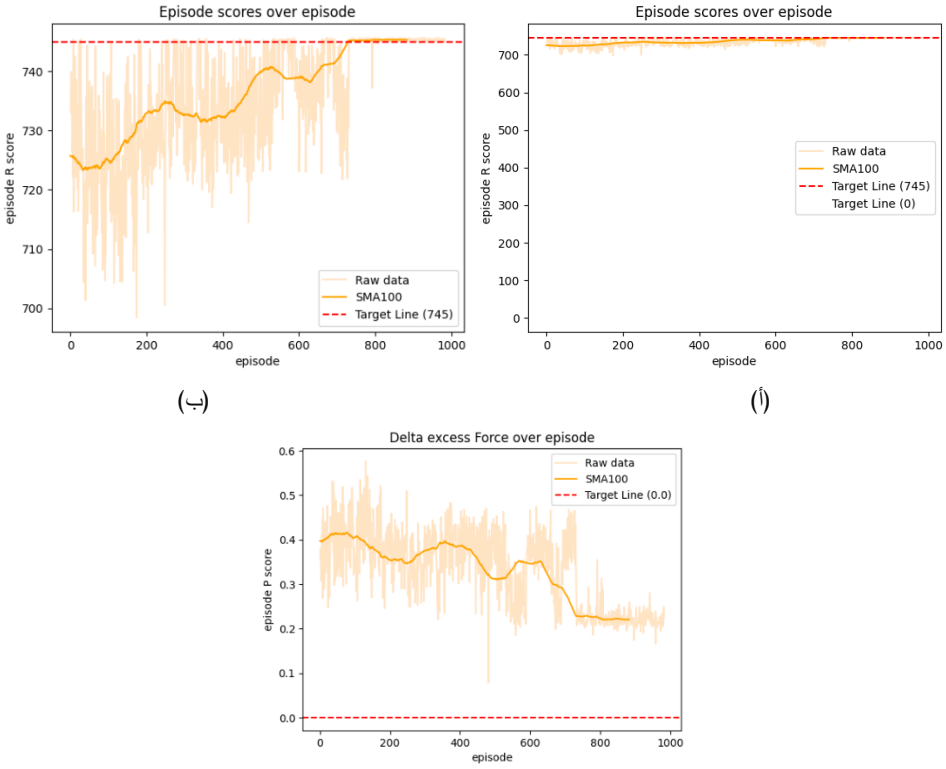
(ب) (أ)

الشكل 6–21 نتائج اختبار السياسة الناتجة عن الاختبار الرابع. (أ). SMA للمكفأة التراكمي. (ب). SMA لفرق الضغط التراكمي.



الشكل 6–22 فرق الضغط المطبق في كل خطوة عند اختبار حلقة واحدة.

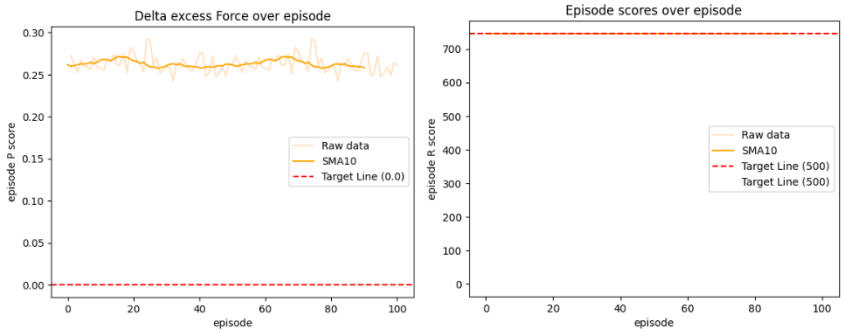
5) **نتائج الاختبار الخامس:** يبين الشكل (23 – 6 (أ)(ب)) استقرار النموذج بعد ما يقارب 1000 حلقة تدريب إضافية، أما فرق الضغط التراكمي الموضح في الشكل (23–6 (ج)) فاستقر وخفض تقريبا إلى قيمة 0.2 نيوتن. وكما يوضح الشكل (24–6) عند اختبار السياسة في هذه المرحلة حققت Success rate test = 100% وفرق الضغط التراكمي تقارب مع 0.25 نيوتن. وكانت النتيجة النهائية مستقرة بالنسبة للحلقة الواحدة، ولا يتم إفلات العبوة بعد إنهاء الخطوات جميعها، أي يبقى ممسكا بها مع محاولة عدم تجاوز فرق الضغط لقيم 0.2 نيوتن كما يوضح الشكل (25–6). وهنا تبين أن النتائج مرضية كما ونوعا لذا تم اعتماد هذه السياسة كسياسة نهائية أجريت عليها الاختبارات اللاحقة.



(ب) (أ)

(ج)

الشكل 23–6 منحنيات التدريب للاختبار الخامس من مرحلة البناء. (أ). منحني المكافأة التراكمية مع 1000 حلقة إضافية مع حد نجاح 745. (ب) هي المنحني أ ولكن مكبر. (ج) منحني فرق الضغط التراكمي في 1000 حلقة إضافية.



(ب) (أ)

الشكل 24–6 نتائج اختبار السياسة الناتجة عن الاختبار الخامس. (أ). المتوسط المتحرك للمكافأة التراكمية. (ب). فرق الضغط التراكمي كمتوسط.



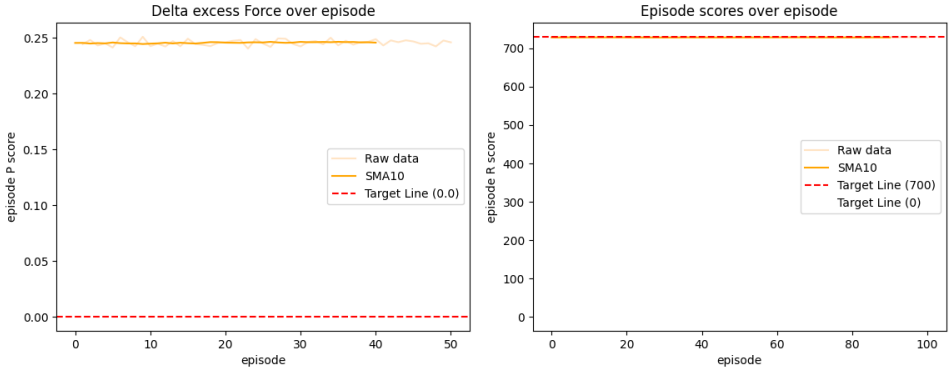
الشكل 25–6 فرق الضغط المطبق عند اختبار حلقة واحدة.

6.3.2.نتائج اختبارات تقييم الصلابة (المرحلة الثانية):

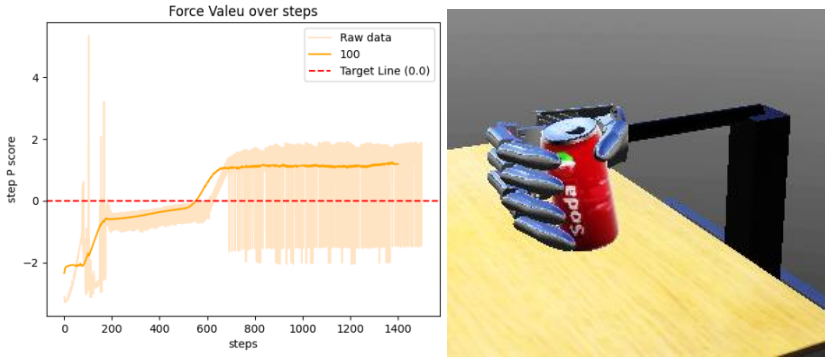
بعد التثبيت على السياسة السابقة كسياسة نهائية تم إجراء عدة تجارب بهدف اختبار القدرة على التعميم ومدى صلابة السياسة، وفي البداية اختبرت السياسة على ذات الإعدادات في التدريب حيث حققت نسب نجاح 100%. ثم في تجارب تعديل الموضع بالإزاحة 5 سم أفقيا للعبوة المعدنية كما يوضح الشكل (26–6) للنتائج الكمية والنوعية تمكن الوكيل مع هذه السياسة من الإمساك وباستقرار ونجاح أكبر وهذا أمر منطقي لأن هذا الموضع أنسب للإمساك كما تبين نتائج المحاكاة حيث حقق نسبة نجاح 100% وفرق ضغط تراكمي 0.2 نيوتن.

عند التجربة الأخرى بتعديل موضع الجسم أفقيا مع إزالة الطاولة، لم يستطع الإمساك وفشل حيث حقق نسبة نجاح 0%، أما فرق الضغط المطبق فقد وصل متوسطه في بعض الحلقات لقيم 3 نيوتن طبعا بالقيمة المطلقة ما يدل على ان معظم الحالات كان الضغط المولد 0 وأن السياسة هنا أصبح غير قادرة على تحقيق القبضة قبل سقوط الجسم وهو ما دل على أن السياسة أصبحت أكثر بطئا من حالة حساسات اللمس وذلك لأنها أصبحت أكثر حذرا، فالآن لا يتعين عليها التقاط الجسم بشكل أعمى، بل يجب أن تلتقط العبوة بأقل قيمة إضافية ممكنة للضغط وكما يبين الشكل (27–6) لأحد الحلقات، فقيم فرق الضغط جميعها أدنى من الحد المطلوب لأنه لم يحقق الإمساك.

مع إعادة ذات التجربة السابقة ولكن برفع العلبة المعدنية 10 سم نحو الأعلى ليتثنى له التقاطها، رغم أنه لازال يفشل في الوصول لمكافأة تراكمية 745 والتي عينت كحد نجاح إلا أنه تمكن في بعض الحلقات من تحقيق مكافأة تصل لـ333 بالإضافة للكثير من الحلقات الفاشلة التي حققت قيم أقل من ذلك بكثير، ويؤكد هذا أن إضافة القوة هنا ساهمت في تقليل سرعة الطرف بهدف أمثلة القوة المطبقة، على عكس الحالة السابقة مع حساسات اللمس والتي حققت في تجربة شبيه نسبة نجاح 100% وتمكن من التقاط الطرف أثناء السقوط في أجزاء من الثانية لأنه كان يسعى فقط لتوليد قبضة أسطوانية ولا يراقب أي عامل ثاني، وهذه النتيجة طبيعية إذ أنه من المنطقي مع إضافة قيود أكثر أن يصبح النموذج أبطأ وأكثر حذرا. يعرض الشكل (28–6)، نتائج هذه التجارب وكما يعرض الشكل (28–6 (د)) النتائج لقيم فرق الضغط لثلاث حلقات متتالة الأولى منهم (أول 1500 خطوة) حققت مكافأة 333 وتمكنت من الإمساك، أما الحلقتين التاليتين لم يتمكنا من تحقيق الضغط المطلوب وفشلت في الإمساك.

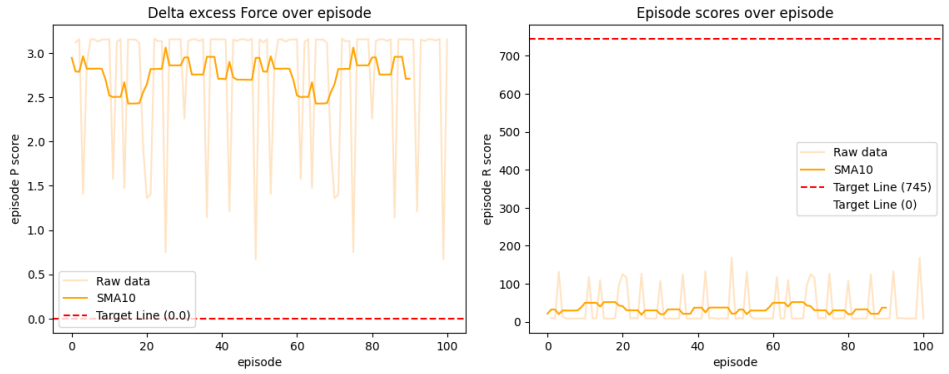


(ب) (أ)

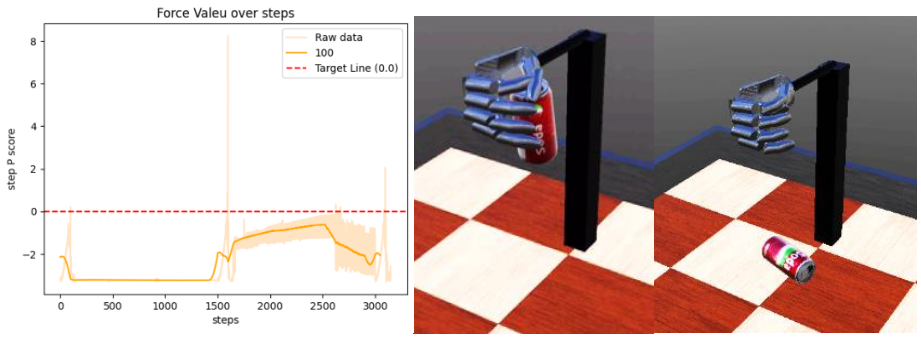


(د) (ج)

الشكل 26–6 تجربة الإزاحة. (أ). SMA للمكافأة التراكمية. (ب). SAM لفرق الضغط التراكمي، (ج). النتيجة العينية، (د) النتيجة لحلقة اختبار واحدة.

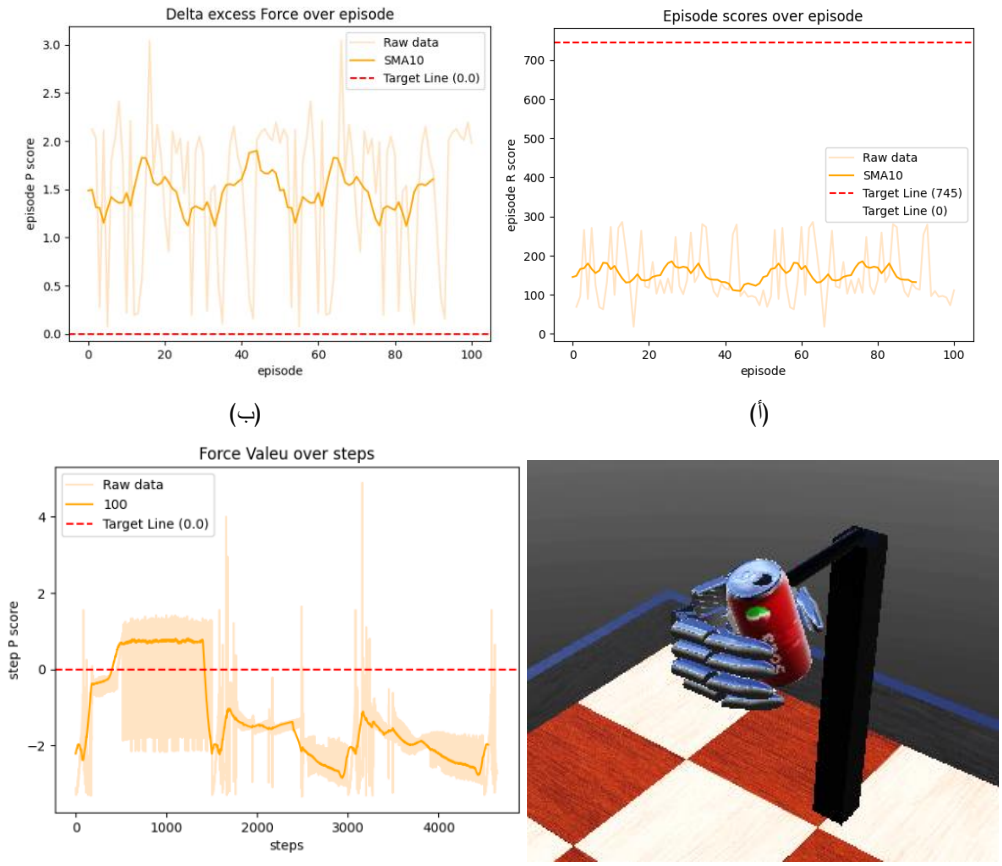


(ب) (أ)



(د) (ج)

الشكل 27–6 تجربة الإزاحة، ومع إزالة الطاولة. (أ). المكافأة التراكمية. (ب). المتوسط المتحرك عبر الحلقات لفرق الضغط التراكمي كمتوسط لكل حلقة، (ج). النتيجة العينية، (د) النتيجة لحلقة اختبار واحدة.



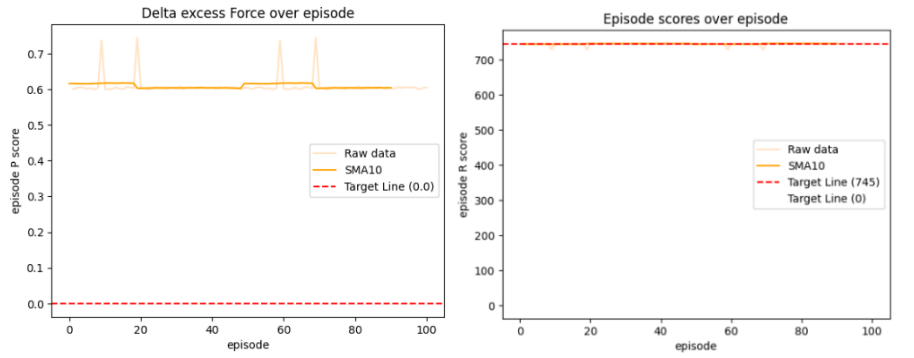
(ب) (أ)

(د) (ج)

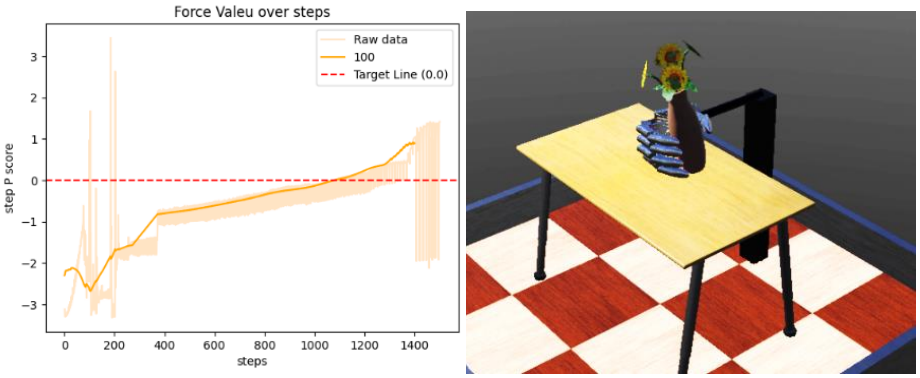
الشكل 28–6 إعادة التجربة السابقة مع رفع العبوة 10 سم للأعلى (أ). المكافأة التراكمية. (ب). SAM لفرق الضغط التراكمي، (ج). النتيجة العينية، (د) النتيجة لثلاث حلقات اختبار.

عند التجريب على أجسام مختلفة مع جعلها بذات الوزن، حيث درست حالة زجاجة مزاحة 5 سم أو أقل، ومزهرية مزاحة 2 سم عن الموضع الأولي للتدريب، مع المزهرية وكما يبين الشكل (6-29) بالنسبة للتقييم حسب تابع المكافأة التراكمية مع حد نجاح 745 حقق النموذج نسبة نجاح 92% وأما القيم التي خالفت حد النجاح 745 لم تكن أقل من 725 وهي أيضا يمكن اعتبارها ناجحة بشكل عام عند مراقبة النتائج العينية ضمن المحاكاة، أما فرق الضغط المطبق وبما أن القبضة هنا أضيق أخذ الوكيل وقتا حتى تمكن من توليد الضغط وعندما قبض وبدأ بالاستقرار كانت الخطوات قد انتهت في الحلقة، وبالتالي كان بحاجة خطوات أكبر لإتمام هذه المهمة بنجاح أفضل.

ومع الزجاجة في الشكل (6-30)، وإزاحة أقل من 5 سم نجح بنسبة 100% إذ أن القطر الأكبر للجسم ولد قبضة كبيرة تمكن من الوصول وتوليد الضغط فورا. أما عند الإزاحة أكبر من 5 سم فشل الوكيل بتشكيل قبضات ناجحة الشكل (6-31) وحقق 0% ولكن اللافت أنه كان مستقر بشكل جيد. ويعرض الشكل (6-32) منحني لحلقتين الأولى مع إزاحة أقل من 5 سم والآخر أكبر من 5 سم. ويلاحظ أن الضغط يختلف كثيرا بحسب شكل القبضة ومكان الجسم الملتقط وتوزع الكثافة رغم أنها متجانسة إلا أنه يتأثر بشكل العبوة (مثلا في العبوة المعدنية تتوزع الكثافة على شكل أسطواني متناظر أكثر من المزهرية والزجاجة)، وتدل هذه النتائج أن النظام مع إضافة حساسات القوة أصبح أكثر حساسية للتغيرات من وجود حساسات لمس، ورغم ذلك تكون القبضات هنا أكثر جودة وموثوقية.

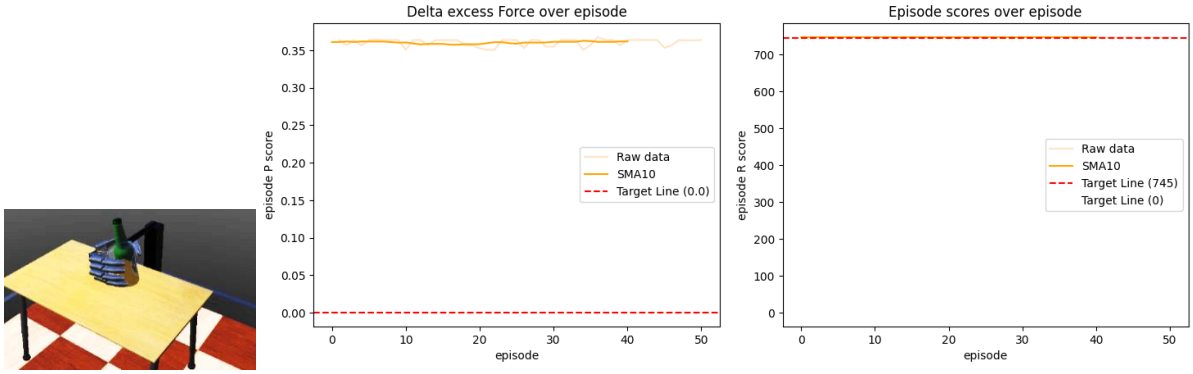


(ب) (أ)



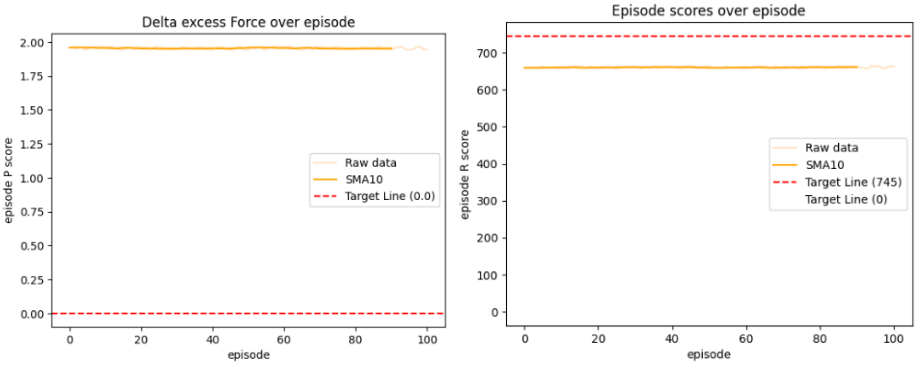
(د) (ج)

الشكل 29–6 التجربة مع مزهرية مزاحة، (أ) منحني المكافآت التراكمية، (ب) منحني فرق الضغط كمتوسط بالقيمة المطلقة، (ج) نتائج عينية (د) أحد الحلقات



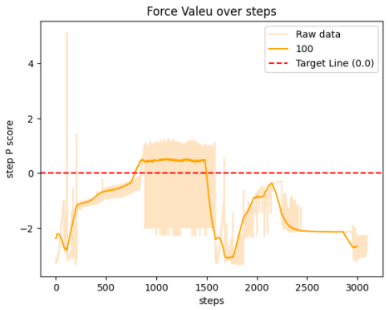
(ج) (ب) (أ)

الشكل 30–6 زجاجة مع إزاحة أقل من 5 سم، (أ) منحني المكافآت التراكمية، (ب) منحني فرق الضغط كمتوسط بالقيمة المطلقة، (ج) نتائج عينية.



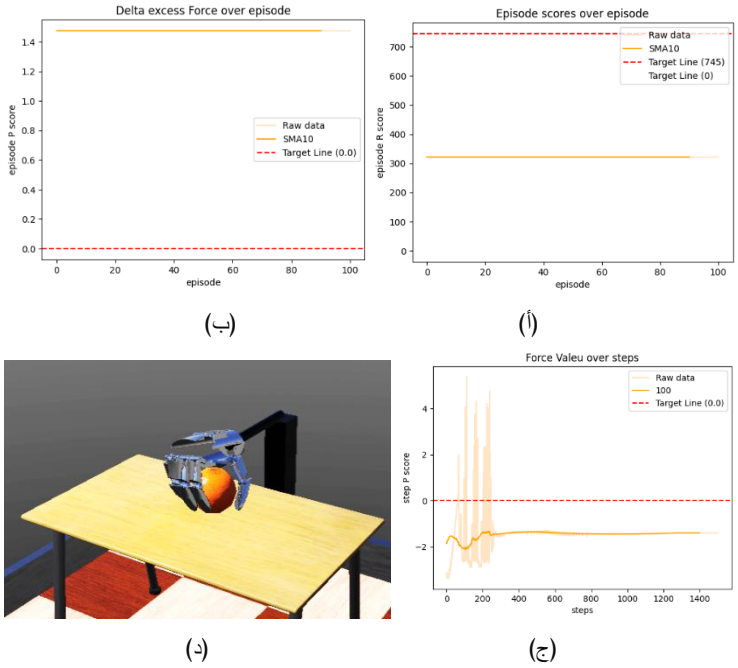
(ب) (أ)

الشكل 31–6 إزاحة أكبر من 5 سم، (أ) منحني المكافآت التراكمية، (ب) منحني فرق الضغط كمتوسط بالقيمة المطلقة.



الشكل 32–6 حلقتين مع إزاحة أقل من 5 (ناجحة) ويليها مع إزاحة أكبر من 5 سم (حلقة فاشلة).

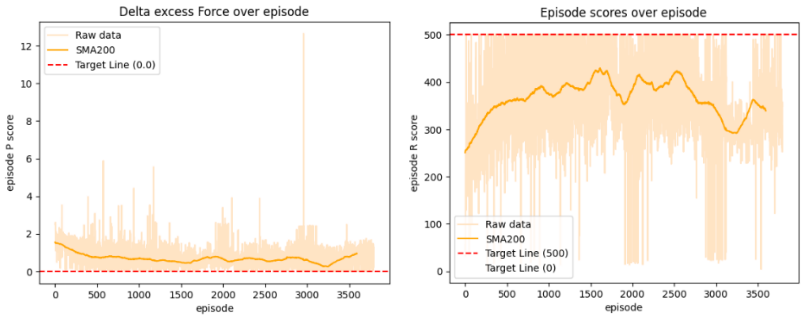
عند تجربة شكل قبضة مختلف (كروي) وتَوضع مختلف للطرف مع برتقالة بوزن 300 غرام استطاعت السياسة أن تكون مستقرة ولكنها لم تولد أي قبضة ناجحة وفشل النموذج بتحقيق قبضة وضغط مناسبين كما يوضح الشكل (6-33 (أ) و(ب)) أن أعلى مكافأة تراكمية تمكن النموذج من تحقيقها هي 330 تقريبا وكانت قيمة مستقرة وأما المتوسط المتحرك لفرق الضغط التراكمي بالقيمة المطلقة استقر على قيمة 1.4N تقريبا عند عرض منحني الضغط المطبق عن القيمة المرغوبة لحلقة واحدة كما يوضح الشكل (6-33 (ج)) تكون دون القيمة الهدف رغم تشكيل القبضة إلا أن القوة غير كافية للرفع.



الشكل 33–6 البرتقالة (قبضة ووزن وتوضع كف) مختلفين (أ) المكافأة التراكمية، (ب) فرق الضغط التراكمي، (ج) فرق الضغط لحلقة واحدة (ج)، نتائج عينية.

6.3.3.نتائج اختبارات تعلم المناهج (المرحلة الثالثة):

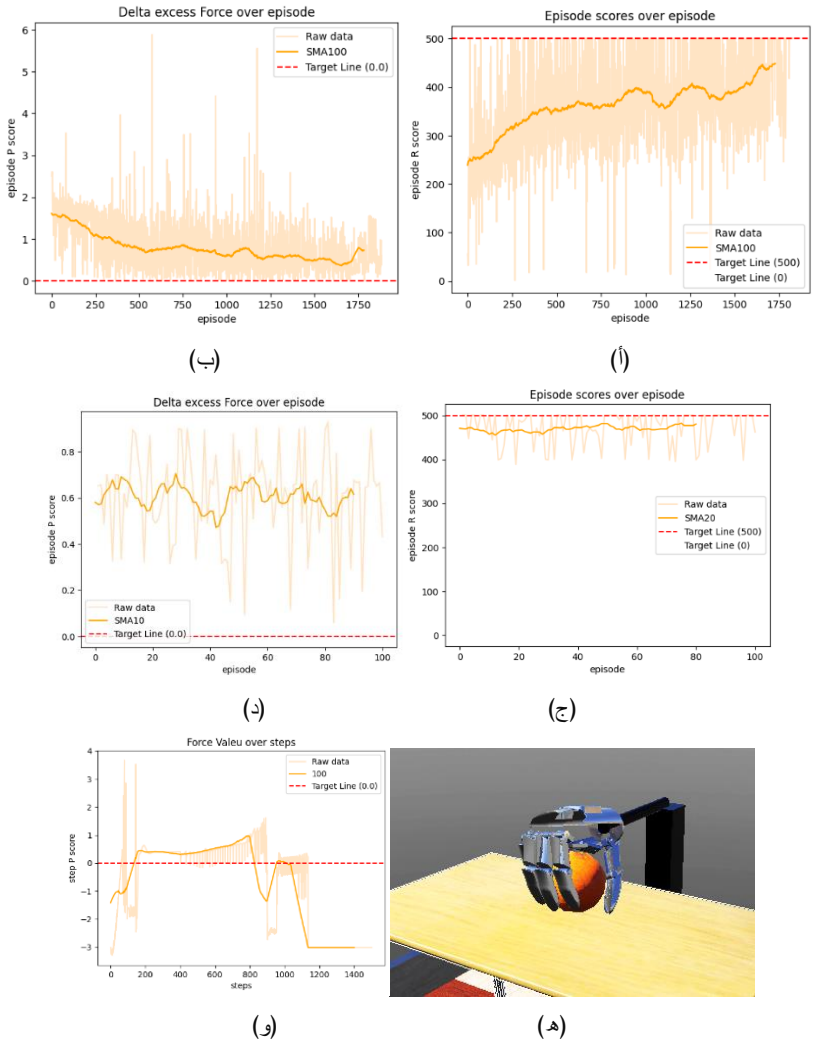
بينت التجربة الأخيرة السابقة حدود النموذج إذ لوحظ عدم تمكن النموذج من التعامل مع قبضات مختلفة على عكس حالة اللمس والتي كانت أكثر عمومية، ولتضمين هذه الحالة في خبرات النموذج تم تطبيق تعلم المناهج بمتابعة تدريب النموذج السابق 5000 حلقة جديدة دفعة واحدة مع تعيين حد النجاح على 500، ولوحظ أثناء التدريب كما يبين الشكل (6–34) ارتفاع ممتاز في قيمة المكافأة لكن بعد الحلقة 2500 عاد لينخفض مرة أخرى، أما قيمة المتوسط الحسابي عبر الخطوات بالقيمة المطلقة للحلقة أيضا بعد أن انخفض عاد للصعود مرة أخرى، لذا تم إيقاف هذا التدريب يدويا، وإعادة التدريب على هذه المهمة بطريقة مدرجة.



(ب) (أ)

الشكل 34-6 تعلم المناهج لدفعة واحدة.

عند إعادة التدريب تم البدأ مع 1800 حلقة لوحظ كما تبين الأشكال في الشكل (35-6) تحسن للنتائج في التدريب وأما في الاختبار نجح النموذج بنسبة 53% ولوحظ أن الحلقات التي كانت أقل من 500 تملك مكافآت تراكمية جميعها فوق 450 لذا تمت متابعة التدريب لمدة 700 حلقة إضافية وكما يوضح الشكل (36-6) استطاع النموذج الاستقرار وتحقيق نسبة نجاح في الإمساك 100% كما كان فرق الضغط المطبق على البرتقالة تقريبا معدوم أو أكثر قليلا من 0 وهي حدود مقبولة ويمكن القول إن هذه الحالة هي أفضل قبضة ولدها النموذج حتى الآن.

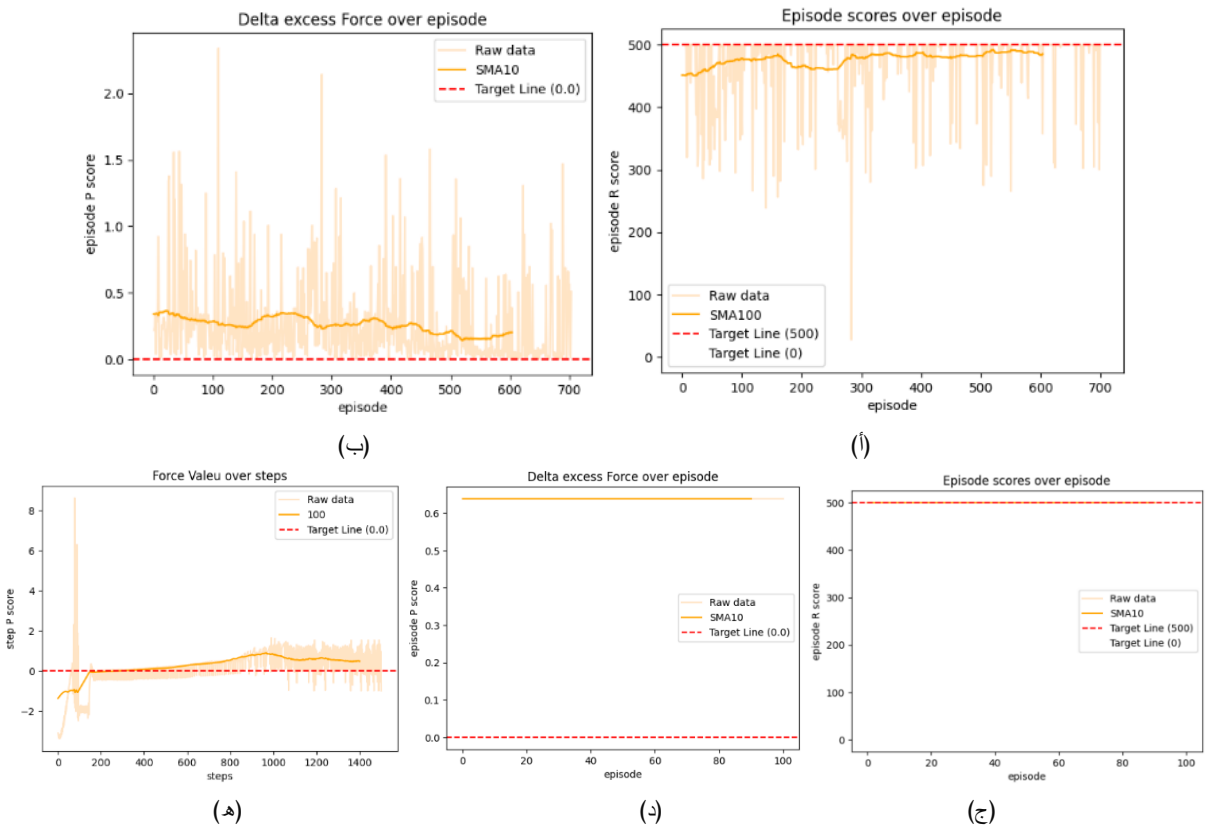


(ب) (أ)

(د) (ج)

(و) (هـ)

الشكل 35-6 تعلم المناهج على مرحلتين مرحلة 1، (أ) المكافأة التراكمية، (ب) فرق الضغط التراكمي، (ج) (د) نتائج الاختبار، (هـ) نتائج عينية، (د) فرق الضغط لحلقة واحدة.

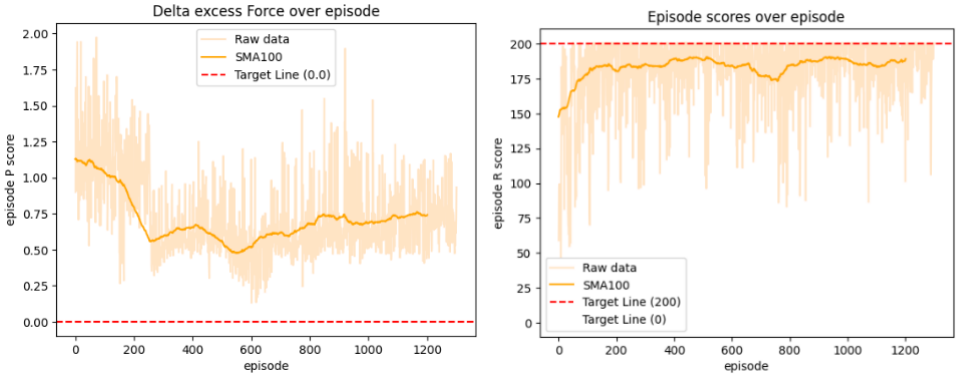


الشكل 36–6 تعلم مناهج مرحلة 2 (أ) المكافأة التراكمية، (ب) فرق الضغط التراكمي بالقيمة المطلقة، (ج) (د) نتائج الاختبار، (هـ) فرق الضغط لحلقة واحدة.

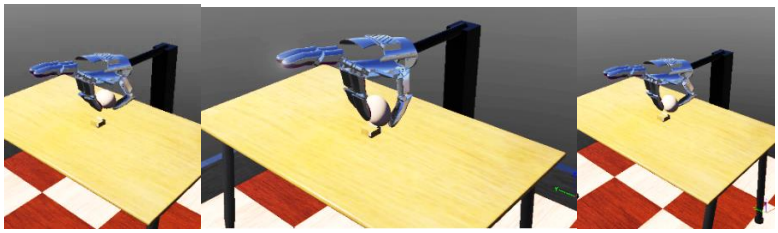
6.3.4.نتائج اختبار قوة النظام المقترح على إنشاء قبضات مخصصة (المرحلة الرابعة):

في هذه المرحلة تم التدريب بدأ من الصفر على مهمة جديدة كليا كما ذكر في الأعلى، بينت النتائج أثناء التدريب أن النموذج كان قادرا على خفض الضغط ورفع المكافأة التراكمية لكنه أيضا كان من الصعب جدا عليه أن يستقر بشكل كامل، ويعذى ذلك بشكل أساسي لشكل الجسم الملتقط وهو بيضة الدجاج البيضوية الشكل، استطاع النموذج بعد تدريب دام 1300 حلقة لكل منها 2000 خطوة وتعيين حد النجاح على قيمة 200 رغم أنه استطاع تحقيق مكافأة تراكمية اعظمية وصلت حتى 200 في الكثير من الحالات وفرق ضغط تراكمي بقيمة 0.75 نيوتن، إلا أنه وفي الاختبار اعتبرت قيمة 163 قيمة النجاح وحقق نسبة نجاح عند الاختبار قدرها 88% علما أن القيم التي تبعد عن هذه النسبة تختلف عنها بمقدار صغير جدا وجميعها تقع فوق 150.

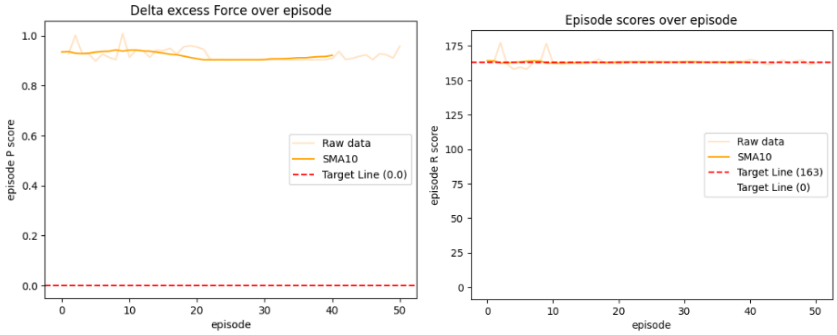
تعرض نتائج هذا الاختبار في الشكل (37–6)، ويتبين أيضا أن فرق الضغط التراكمي المأخوذ كمتوسط حسابي بالقيمة المطلقة عبر الحلقات حقق في الاختبار قيمة 0.9 نيوتن وأثناء عرض قيم الضغط لحلقة اختبار ناجحة واحدة تبين أن النموذج كان شبه مستقر وفي الحلقات الناجحة يطبق فرق ضغط زائد تقريبا 2 نيوتن وهو ما يعتبر ضمن النطاق المقبول كما عرض(Altuntas et al, 2008) (Zhang et al, 2015)، كا أنه في بداية التلامس يطبق ضغط أكبر من ذلك لكنه أيضا ضمن الحدود المقبولة.



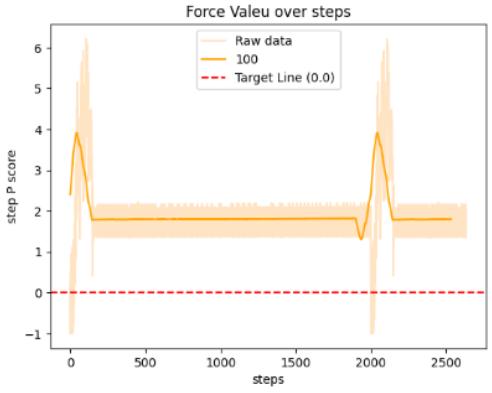
(ب) (أ)



(ج)



(هـ) (د)

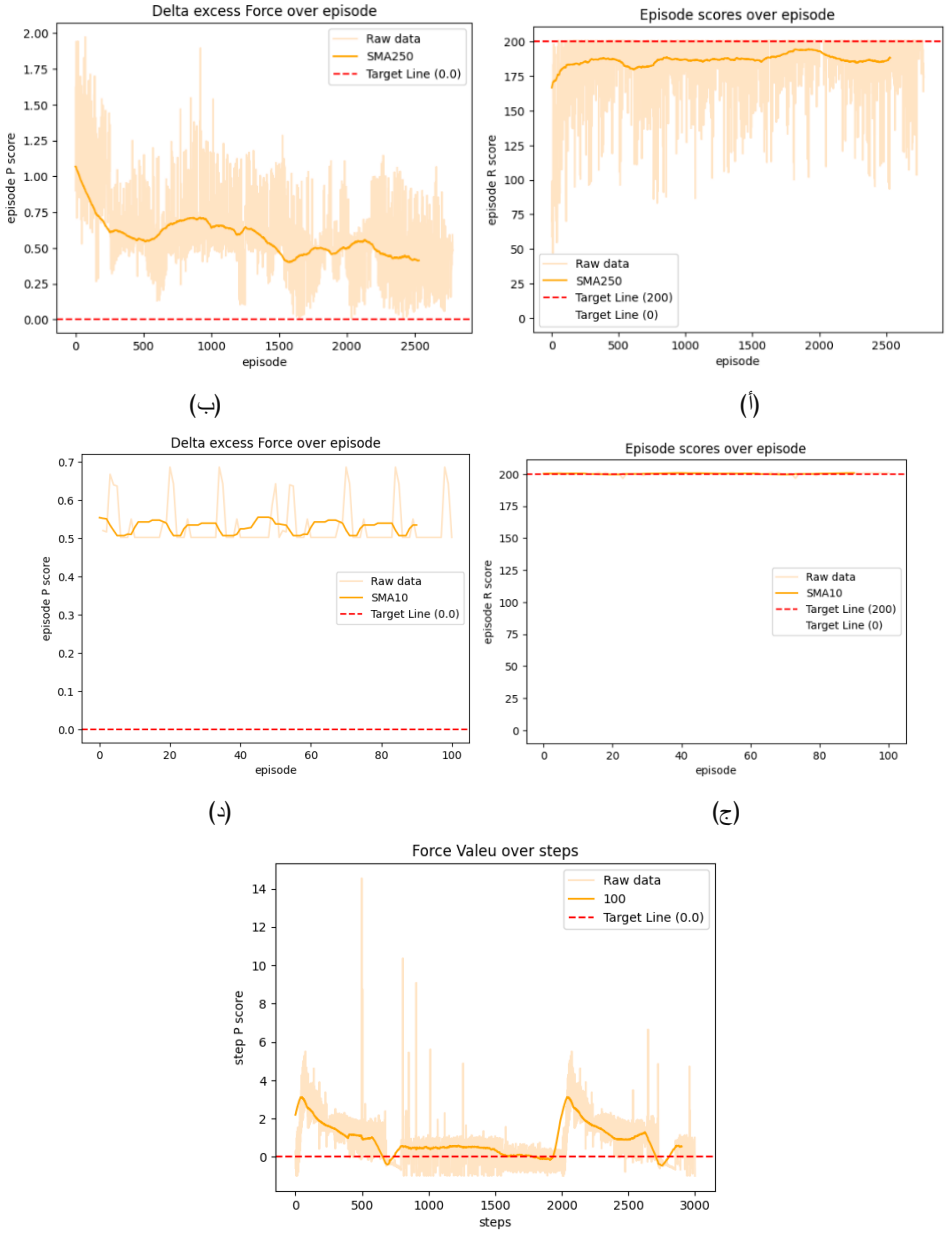


(و)

الشكل 37–6 تجربة إمساك بيضة دجاجا بثلاث أصابع. (أ)، المكافأة التراكمية، (ب) فرق الضغط التراكمي، (ج) نتائج عينية، (د) (هـ) نتائج الاختبار، (و) نتائج حلقة وبداية الحلقة التالية.

مع إزاحة بيضة الدجاج 2 سم أفقيا أبدى نتائج أفضل كقيمة عظمى عائدة لأن شكل القبضة كان أفضل للإمساك حيث حقق 500 ببعض الحالات، ولكن ازداد الضغط المطبق. وبالمجمل حقق نسبة 58% تقريبا كنسبة نجاح مع اعتبار حد النجاح 200.

ولأمثلة السياسة السابقة تم متابعة علمية التدريب للنموذج دون إزاحة وذلك رغبة في أن يستقر النموذج أكثر مع متاعبة التدريب مايقارب 1000 حلقة إضافية 2000 خطوة لكل منها وحد نجاح 200 تحسنت النتائج من 88% إلى قيمة 96% مع خفض فرق الضغط التراكمي من قيمة 0.9 نيوتن إلى 0.6 نيوتن ورغم ذلك لم يكن النموذج مستقر بشكل كامل ويعود السبب لذلك في أن الهيكل البيضوي المصمم صعب أن يولد قبضات مستقرة بثلاث أصابع لها التركيب الهندسي للطرف المختار ويعتقد أن للشكل الهندسي لبيضة الدجاج المصممة في هذه الدراسة تأثير واضح على المهمة وأن الاختبار مع بيضة دجاج بشكل هندسي أكثر بساطة من الممكن أن يعطي نتائج أفضل. يوضح الشكل (38–6) نتائج هذه التجربة.



الشكل 38–6 التجربة لإمساك بيضة دجاج مزاحة 2 سم، (أ)، المكافأة التراكمية، (ب) فرق الضغط عبر الحلقات، (ج) نتائج عينية، (د) (هـ) نتائج الاختبار، (و) نتائج حلقة وبداية الحلقة التالية.

6.4. مناقشة النتائج ومقارنتها فيما بينها:

أظهرت نتائج البحث فعالية واضحة في تصميم سياسة تحكم قوية تعتمد على التعلم المعزز العميق، قادرة على تعميم المهام وتجاوز التحديات المرتبطة بتعدد أشكال الأجسام ووضعياتها. كما أظهرت النتائج اختلافا ملحوظا في الأداء للنماذج المدربة مع اختلاف نوعي الحساسات، في هذه الفقرة سيتم عرض ملخصا للنتائج ومصفوفة الأداء لكل التجارب والاختبارات السابقة لكل مرحلة على حدا ومناقشتها ومقارنتها فيما بينها ومع الدراسات السابقة.

6.4.1. ملخص نتائج حالة حساسات اللمس (الحالة الأولى) ومناقشتها:

يُمثل العمود الأول في الجدول (2-6) التجربة، وأما العمود الثاني فهو فضاء الإجراءات المعتمد والعمود الثالث يعبر عن نوع هيكل المكافأة الذي يتم اختباره، في حين يعبر العمود الرابع عن عدد الحلقات الأعظمي لكل تجربة (علما أن هذا العدد قد يكون عمليا أقل من ذلك إذا تم إيقاف الحلقات يدويا أو تم إكمال المهمة، وتعتبر المهمة مكتملة SOLVED إذا كان المتوسط الحسابي للمكافآت التراكمية لآخر 350 حلقة منها يساوي قيمة حدية تعين بالتجربة والخطأ لكل اختبار ومرحلة دراسة، يمثل العمود الخامس عدد الخطوات في كل حلقة، في حين يعبر العمود السادس عن عدد الخطوات الكلية في كل تجربة وينتج هذا الرقم من جداء عدد الحلقات الكلية بعدد الخطوات ضمن كل حلقة، يملك العمود السابع شرط نجاح الحلقة Is-Done إذ تعتبر الحلقة ناجحة إذا كانت قيم المكافآت التراكمية التي تم جمعها عبر الخطوات أكبر من حد معين يتم تعينه لكل حالة وتجربة عن طريق التجربة والخطأ، ثم يأتي شرط إكمال المهمة أو نجاحها في العمود الثامن والذي يملك شرط الإكمال وتعتبر المهمة مكتملة وتم حلها Solved إذا كان كما ذكرنا المتوسط الحسابي للمكافآت التراكمية لآخر 350 حلقة منها يساوي شرط نجاح الحلقة وبمعنى آخر (إذا كانت الغالبية العظمة لآخر 350 حلقة ناجحة سيتم اعتبار المهمة مكتملة)، يعرض العمود التاسع من الجدول التقارب مع القيمة المرغوبة وبجواره الزمن الذي بدأ عنده النموذج بالتقارب وهنا يحسب الزمن بالعلاقة (عدد الحلقات قبل أن يبدأ التقارب × عدد الخطوات × (زمن الخطوة الواحدة في الحلقة وهو يساوي زمن خطوة المحاكاة 0.032 ثانية)) مقسما على 3600 للتحويل إلى ساعات، بعد التقارب يتم التدريب عدة حلقات إضافية بهدف الاستقرار ويعرض في العمود قبل الأخير وقت التدريب الكلي ضمن المحاكاة، وجواره يعرض أخيرا معدل النجاح لكل تجربة تم اختبارها على مدى 100 حلقة اختبار في الغالب.

في تجارب البناء والتدريب (المرحلة 1)، والتي تمثل مجموعة من التجارب المكثفة أجريت لتحديد الآلية الأنسب من حيث نوع تابع المكافأة وفضاء الإجراءات الواجب اتخاذه وتعين الحدود التي تعتبر عندها الحلقات ناجحة والمهمة مكتملة، وفي هذه المرحلة تم إثبات فعالية دور هيكل المكافأة وإعدادات التدريب وأثرها الكبير على النتائج، حيث تم التأكيد على أن التصميم المناسب لتابع المكافأة هو العامل الأساسي في نجاح نموذج التعلم المعزز. وتبين أخيرا بعد جملة من التجارب أن استخدام هيكل المكافأة B مع الإجراء B، ضمن إعدادات مدروسة (عدد حلقات وخطوات

كاف)، أتاح للنموذج التعلم بسرعة وكفاءة، محققا استقرارا في الأداء في أقل من 400 حلقة (أي ما يقارب ساعتين ونصف حتى تدرب وبدأ بالتقارب بعد أقل من ساعة تدريب). يشير ذلك إلى أن التوازن بين المكافآت المرحلية والتراكمية النهائية في الهيكل المقترح عزز التعلم التدريجي السليم وقاد إلى سلوك أمثل.

في تجارب التعميم (المرحلة2)، وهي اختبار قدرة السياسة على التعميم (اختبار صلابتها)، تبين أن السياسة المدربة أظهرت صلابة قوية تجاه تغييرات متعددة في البيئة: كتغير نوع الجسم (قارورة ماء أو زجاج)، وتغير موضع الجسم (إزاحة أفقية)، أو حتى تغير الشكل بالكامل (برتقالة، تفاحة). نجاح السياسة في معظم هذه السيناريوهات بنسبة 100% يؤكد قدرتها على التعميم ضمن حدود معينة. كما أكد نجاح الوكيل في إلتقاط جسم أثناء سقوطه سرعة الوكيل في تنفيذ القبضة المطلوبة منه في أجزاء من الثانية ثم الحفاظ عليها باقي خطوات الحلقة، إلا أن فشل النموذج في التعامل مع إزاحة أفقية كبيرة للجسم الكروي كشف عن قصور في مدى تعميم السياسة على أوضاع لم تظهر في بيانات التدريب. وهذه النتيجة كانت متوقعة، كون السياسة تعتمد بشكل كبير على البيانات التي تم تدريبها عليها، كما أن شكل الجسم وقربه من الإبهام بدلا من الأصابع الأربع مثل تحديا هندسيا لم يكن مأخوذا بعين الاعتبار ضمن السياسة المدربة.

في تجارب تعلم المناهج (المرحلة3)، وهي دراسة أثر تعلم المناهج على أداء النموذج، كان لإدخال تعلم المناهج خطوة استراتيجية سمحت بتوسيع نطاق الخبرة لدى النموذج. وفعلا، أدى إدخال الجسم المزاح تدريجيا في سياق التعلم إلى انتقال النموذج من فشل تام 0% إلى أداء مثالي 100%. وذلك بوقت قصير نسبيا إذ حصل التقارب بعد مايقارب ساعة تدريب إضافية واستقر التدريب بعد ساعتين تقريبا وهو ما لم يكن ليحققه بالتدريب بدأ من الصفر على هذه المهمة. هذه النتيجة توضح أن تعلم المناهج يمكن أن يكون أداة فعالة لتجاوز حدود البيانات الأولية، إذ يمكّن النموذج من التعامل مع تحديات تدريجية قد لا ينجح أمامها مباشرة، عن طريق تنظيم إدخال البيانات والمهام على شكل تدريجي بدأ من الأسهل.

في تجارب اختبار قوة النظام المقترح على إنشاء قبضات مخصصة مستقبلا (المرحلة 4)، أظهرت التجارب أن المنهجية وهيكل المكافأة المقترح قادر على دعم تعلم فعال حتى عند البدء من الصفر لتعلم إمساك شكل جديد كليا (مطرقة). ورغم أن الشكل الجديد تطلب زيادة في عدد الخطوات وحد المكافأة للوصول إلى أداء ناجح، فإن السياسة استطاعت التكيف تدريجيا مع الوضع الجديد وحققت نجاحا كاملا في ظروف متعددة (جسم متحرك، إزاحة أفقية، ورفع حد النجاح). هذا يبين أن هيكل المكافأة لم يكن مرتبطا فقط بالجسم الأولي (الأسطواني)، بل كان عامّا بما فيه الكفاية ليسمح بتعلم أنماط قبض جديدة عند توفير التدريب المناسب. وبالتالي، فإن هذا الهيكل يعد مرشحا واعدا لاستخدامه في أنظمة الأطراف الاصطناعية، حيث يمكن تدريب النموذج لاحقا على أجسام خاصة بكل مستخدم حسب نمط حياته، مما ينتج أطراف تفاعلية.

الجدول 2-6 ملخص التجارب ومصفوفة الأداء لحالة حساسات اللمس Performance metrics

	التجارب	إجراء	مكافأة	حلقات	خطوات	الوقت الكلي	Is-done	Slved	التقارب	وقت التقارب (ساعة)	زمن التدريب (الساعة)	Success rate
مرحلة 1	1	A	A	5000	700	3.5M	CR ⩾ 180	CRAL⩾180	×	-	>30	-
	2	A	A	5000	2000	10M	CR ⩾ 180	CRAL⩾180	×	-	>88	-
	3	A	B	5000	2000	10M	CR⩾ 1800	CRAL⩾1800	✓	≈ 8	>50	-
	4	B	B	5000	700	3.5M	CR ⩾ 180	CRAL⩾180	✓	≈ 1	>3	100%
مرحلة 2	الكل عدا 4 و8	اختبرت سياسة اتخاذ القرار الناتجة عن الاختبار 4 في المرحلة 1 مع تجربة قبضات ومواضع مختلفة دون تغير إعدادات السياسة										100%
	4	أزيحت العلبة المعدنية أفقيا مع إزالة الطاولة، تمكن الطرف من إلتقاطها قبل السقوط في أغلب الحالات.										91%
	8	إزاحة التفاحة أكثر من 5 سم أفقيا.										0%
مرحلة 3	تدريب +	B	B	500+	700	350+ ألف	CR ⩾ 180	CRAL⩾180	✓	≈ 1	>2	100%
	تعميم	تعميم سياسة اتخاذ القرار السابقة على جسم جديد كليا.										61%
مرحلة 4	1	B	B	5000	700	3.5M	CR ⩾ 180	CRAL⩾180	×	-	>6	-
	2	B	B	500	1500	750 ألف	CR ⩾ 350	CRAL⩾350	✓	≈ 3	>6	100%
	3	B	B	400+	1500	600+ ألف	CR ⩾ 600	CRAL⩾600	✓	≈ 2	>5	100%
	4	التعميم على جسم مزاح.										100%

6.4.2. ملخص نتائج حالة حساسات القوة (الحالة الثانية) ومناقشتها:

في الجدول (3-6) يضاف فقط العمود P والذي يعبر عن فرق الضغط عبر الخطوات لكل حلقة كمتوسط حسابي بالقيمة المطلقة (أطلق عليه اسم فرق الضغط التراكمي)، وتعرض القيمة التقريبية التي استقر عندها الوكيل.

في تجارب البناء والتدريب (المرحلة 1)، أجريت كما في حالة حساسات اللمس مجموعة من التجارب لتعين المنهجية الأفضل للحصول على سياسة أنسب يمكنها تقديم أفضل الحلول بأنسب وقت ممكن، ولكن في هذه المرحلة وبسبب زيادة كمية المدخلات من الحساسات كون القيم الـ20 للدخل أصبحت مستمرة لم تبدي البنية الأولية التي استعملت في حالة اللمس الثنائي لشبكتي الفاعل الناقد في خوارزمية PPO أي نتائج إيجابية تشجع على إتمام التدريب إذ وحتى عندما بدا النموذج مدربا وناجحا أثناء الاختبار بتقاربه بعد 400 حلقة (5 ساعات تقريبا) من حد النجاح 500 وتمكنه من خفض القيمة المطلقة لمتوسط فرق الضغط بالقيمة المطلقة إلى 1.3 نيوتن، إلا أنه عند الاختبار لم يستطع تحقيق نسب نجاح مرتفعة إذ كانت تقريبا 15% كما يوضح الجدول (3-6)، لذا تم تعديل بنية الشكتين بزيادة العمق لاستخراج أكبر تمثيل ممكن من بيانات الدخل والاستفادة منه، وهي التعديلات التي مكنت من الوصول إلى نسب نجاح 100% واستطاعت تقليل القيمة المطلقة لمتوسط فرق الضغط إلى قيمة نهائية 0.5 نيوتن تقريبا. وبوقت تدريب 6 ساعات تقريبا، ومع زيادة التدريب تحسنت النتائج أكثر إذ رفع حد المكافأة التراكمية الواجب على العميل جمعه حتى 745 مع فرق الضغط المطبق لقيمة 0.2 نيوتن. ولكن أخذ هذا وقت كبير نسبيا وبالتأكيد إن زيادة عبء العمل والدقة المطلوب من النموذج تحقيقها هنا أدت لزيادة الوقت اللازم للحصول على السياسة الأفضل، بل وأيضا سببت بطئ بالنموذج حتى بعد التدريب كما سيوضح الآن.

في تجارب التعميم (المرحلة2)، اختبرت قدرة السياسة على التعميم على حالات جديدة وتغيرات في البيئة، ولوحظ أن السياسة بشكل عام أصبحت أكثر حساسية للتغيرات في توضع الجسم الملتقط والكف، كما أنها أصبح أكثر بطأ

في تنفيذ الأوامر لأن السياسة هنا تنفذ أمر الإلتقاط بحذر تفاديا لتطبيق أي قوة زائدة ويظهر هذا بشكل واضح عند مقارنة إمساك علبة المشروبات المعدنية بعد إزالة الطاولة، ففي حالة حساسات اللمس مجرد بدأ السياسة بالعمل يغلق مباشرة الكف على الجسم وبسرعة فيلتقطه قبل سقوطه ويحقق نجاح بنسبة 100% إلا أنه مع إضافة شرط الضغط يبقى الوكيل حذرا (التجربة 2 من المرحلة 2 في الجدول (3-6)) ويغلق ببطئ أكبر تفاديا لتطبيق قوى زائدة فيؤدي ذلك لسقوط العبوة في كل الحالات قبل إلتقاطها وتحقيق نسبة 0% كنجاح، وكذلك أجريت بعض التجارب التي لم تزكر في الأعلى مع أوزان جديدة وفشل الوكيل معها إذ أن التجريب على أوزان أكبر (1 كغ) كالمطرقة فشل النموذج تماما ومع وزن أصغر (220 غرام) حقق نسب نجاح منخفضة جدا وطبق قوى زائدة لكنها بقيت ضمن الحدود المقبولة.

في تجارب تعلم المناهج (المرحلة 3)، أبدى تعلم المناهج مرة أخرى قدرته على أمثلة النتائج من خلال النهج التدريجي، وحسن النتائج في الإمساك من 0% إلى 53% ثم 100% تدريجا وخفض القيمة المطلقة لمتوسط فرق الضغط حتى 0.5 نيوتن. وبهذا أكد هذا النهج مرة أخرى أهميته، وأثبت مرة أخرى أن النهج المنظم في إضافة الصعوبة تدريجيا على المهام يكون ذو فائدة كبيرة.

في تجارب اختبار قوة النظام المقترح على إنشاء قبضات مخصصة مستقبلا (المرحلة 4)، تمكن الوكيل من تحقيق مهمة جديدة كليا وأكثر دقة بنسبة تصل إلى 96% مع خفض متوسط فرق الضغط التراكمي المطبق من 0.9 إلى 0.6 نيوتن تقريبا في قبضة دقيقة من ثلاث أصابع ويعد نجاح النهج المقدم أمام سيناريو جديد كليا مسألة مهمة للتمكن لاحقا من تدريب النهج المقدم على أشياء خاصة بالمستخدم، كما ذكرنا.

وبالتالي، قدمت إضافة حساسات الضغط قبضة أكثر وثوقية من حيث سلامة الطرف والجسم الملتقط. فمع إضافة حد لقيم الضغط الواجب توليده تمكن الوكيل من توليد سياسة يمكنها رفع الأجسام الأسطوانية ذات الوزن 330 غرام تقريبا أو أقل قليلا بقوة ملائمة. إلا أن إضافة الضغط جعلت السياسة أكثر حساسية وأكثر خصوصية ومقاومة للتغيرات البيئية.

الجدول 3-6 ملخص نتائج الحالة الثانية ومصفوفة الأداء لحالة حساسات القوة Performance metrics يكون الإجراء A والمكافأة C لكل الحالات:

التجارب		الحلقات	الخطوات	الوقت الكلي	Is-done	Solved	تقارب	وقت التقارب (الساعة)	زمن التدريب (ساعة)	P نيوتن	Success rate
المرحلة 1	1	5000	1500	7.5M	CR ⩾ 500	CRAL⩾500	✓	≈ 5	>8	1.75	16%
	2	400+	1500	600+ ألف	CR ⩾ 500	CRAL⩾500	×	-	>5	1.3	18%
	3	500	1500	750 ألف	CR ⩾ 500	CRAL⩾500	✓	≈ 4	>6	0.5	100%
	4	2500+	1500	3.75M	CR ⩾800	CRAL⩾800	×	-	>30	0.4	92% لحد نجاح 730 0% لحد نجاح 800
	5	1000+	1500	1.5M	CR ⩾745	CRAL⩾745	✓	≈ 10	>13	0.2	100%

المرحلة 2	1	تعديل موضع الجسم الملتقط.								100%
	2	تعديل موضع الجسم أفقيا وجعله متحرك								0%
	3	رفع 10 سم تقريبا نحو الأعلى، تمكن من الالتقاط ولكن بأقصى حد للمكافأة 333 وإذا اعتبرت حد نجاح فيكون قد حقق نجاح بنسب منخفضة.								نسبة منخفضة
	4	تجربة أجسام مختلفة بذات الوزن مع انزياحات أصغر من 5سم (الفازة 2 سم) الزجاجة 5 سم وأقل.								نسبة منخفضة
	5	قبضة مختلف وتوضع الكف مختلف: برتقال، بوزن 300 غرام.								0%
المرحلة 3	1	5000	1500	7.5M	CR ⩾500	CRAL⩾500	×	لم يتابع به		-
	2	1800+	1500	+ 2.7 M	CR ⩾500	CRAL⩾500	×	-	24	0.6	53%
	3	700+	1500	+ 1.05M	CR ⩾500	CRAL⩾500	✓	≈ 3	≈ 10	0.5	100%
المرحلة 4	1	2000	2000	4M	CR ⩾200	CRAL⩾200	✓	≈ 3	>23	0.9	88% لحد نجاح 163 0% لحد نجاح 200
	2	اختبار النموذج مع إزاحة								58%
	3	1000+	2000	2M+	CR ⩾200	CRAL⩾200	×	-	>15	0.6	96%

* يمثل solve حد شرط إتمام المهمة، ويمثل is-done حد نجاح الحلقة الواحدة. و CR المكافأة التراكمية للحلقة. أما CRAL هي CR-AVG-LAST 350 هي المتوسط الحسابي للمكافأة التراكمية لآخر 350 حلقة. تشير + إلى تعلم المناهج. أما P فتدل على المتوسط الحسابي بالقيمة المطلقة لقيمة فرق الضغط لكل حلقة (فرق الضغط التراكمي).

6.4.3. مقارنة النتائج فيما بينها:

أخيرا المقارنة بين الحالتين، حيث تمت مقارنة الحالتين بشكل أساسي في المرحلة الرابعة من كل حالة لأنها المرحلة التي تعبر عن المنهجية النهائية المعتمدة والتي تم الحصول والتثبيت عليها للتجارب اللاحقة، وجد أن:

1. **بالنسبة لزمن التعلم ووقت التقارب وزمن التنفيذ بعد التدريب:** فإن إضافة حساسات القوة ساهمت في جعل النموذج أكثر بطأ في التعلم وأداء الحركات بعد التعلم لأنه يكون أكثر حذرا إذ يسعى لتوليد قبضة بأقل ضغط ممكن، وكما ذكر لقد أثبتت تجارب إزالة الطاولة في حالة حساسات اللمس نجح الوكيل بتنفيذ حركة الإمساك بأجزاء من الثانية متمكنا من إلتقاط العبوة المعدنية قبل السقوط على الأرض والحفاظ على هذه الحالة حتى نهاية الحلقة، في حالة حساسات الضغط لم يتمكن من الوصول لوضعية الإمساك الملائمة قبل سقوط العبوة إذ احتاج وقت إضافي أكبر وبالتالي لم ينجح في عملية الالتقاط بهذه الحال، وكذلك كان واضح بمقارنة أوقات التقارب والتدريب أن حساسات اللمس كانت أسرع، وكما ذكرنا يعود هذا إلى حاجة الوكيل للمفاضلة بين القوة والقبضة الملائمة فتابع المكافأة يوجهه لتوليد القبضة الملائمة بحذر بحيث يتجنب تطبيق قوى زائد وكما وضحت مخططات القوة المطبقة في معظم التجارب السابقة أن الوكيل عندما يبدأ بتطبيق الضغط والحصول على التلامس يتراجع في الخطوات اللاحقة لتخفيض هذه القوى المطبقة.
2. **بالنسبة لنجاح الإمساك:** خفض حساس الضغط من الدقة والسرعة بشكل عام، حيث خفضت نسبة نجاح الإمساك العامة للنموذج من 100% في حالة حساسات اللمس إلى 96% في حالة حساسات الضغط رغم أنه حقق 100% في بعض الحالات، ولكن قمنا باعتماد نتائج المرحلة الرابعة بكل تجربة لأنها النتيجة خلاصة التجارب السابقة وهي التي تدل على قوة النهج عند التدريب في السيناريوهات الجديدة كليا، وهذا

منطقي لأن المتطلبات التي تحدد شرط النجاح أصبحت أكبر والبيئة هنا ديناميكية إذ يتغير مركز ثقل البيضة عند محاولة الإمساك وبالتالي أثرت على النجح الذي كان أكثر عمومية في حالة اللمس.

3. **جودة القبضات:** رغم التناقص السابق بالسرعة ونسبة نجاح الإمساك، فإن حساسات الضغط ولّدت قبضات أكثر سلامة ودقة بالنسبة للطرف وللجسم الملتقط وكان مقدار الضغط الزائد المطبق على بيضة الدجاج كقيمة زائدة لا يتجاوز بحسب ما أظهر منحني المتوسط المتحرك لقيم فرق الضغط للخطوة قيمة 4 نيوتن وكما ذكرنا ف إن بيضة الدجاج المصمة يمكنها تحمل ضغط يصل إلى 20 نيوتن قبل أن تنكسر. أما في حالة حساسات اللمس كما أظهرت النتائج العينية أثناء مراقبة القبضات أبدت عدة قبضات قوية أدت لإنهاء الحلقة وابتعاد الجسم الملتقط عن الطرف بشكل مفاجأ ويعبر ذلك في Webots عن حصول خلل ربما كسر في الجسم الملتقط أو قوة زائدة مطبقة عليه أكبر من قوة تحمله.

6.5. المقارنة مع الأعمال السابقة:

بالمقارنة مع الدراسات السابقة ذات الصلة، كما يوضح الجدول (4–6)، استطاع البحث الحالي تقديم تحسن واضح في القدرة على تعليم طرف اصطناعي ذو 10 درجات تحكم توليد حركات الإمساك الملائمة باستخدام التعلم المعزز العميق وردود الفعل اللمسية فقط، وبالتدريب بدأ من الصفر، دون الحاجة لأي بيانات مجمعة مسبقا ماسيساهم في تقليل العبء الحسابي وجهد تجميع البيانات المسبقة.

بداية وبالمقارنة مع دراسات حديثة مثل دراسة (Zhang et al, 2025) التي استخدمت خوارزمية MPO لقابض ثنائي الأصابع مع حساسات لمس ثنائية ومدخلات بصرية وحققت نسبة نجاح 55%، تمكن النهج المقترح باستخدام خوارزمية PPO وكف بخمسة أصابع مع حساسات لمس ثنائية فقط من تحقيق نسبة نجاح بلغت 100%. في حين عندما قام بحث Zhang باستعمال حساسات القوة تمكن من تحسين نسبة النجاح في الإمساك إلى 66% أما في هذا البحث مكنت حساسات القوة من الحصول على دقة 96% كما حققت في بعض الحلات دقة 100%، ويدل ذلك على إمكانية أن تكون إزالة المدخلات البصرية قد ساهمت في تحسين الاستقرار وتقليل الضوضاء، مما يعزز فاعلية نموذج الإمساك المقترح في ظروف معقدة، حيث كما عرض Zhang تقليل جودة الصور مع إضافة مستشعر لمسي يؤدي لزيادة نسب نجاح النماذج، وهو ما يدعم النتائج التي حصل عليها هذا البحث بشكل كبير. حيث ساعد إلغاء الإدخال البصري والاعتماد التام على المدخلات اللمسية بأمثلة النتائج بشكل واضح، وكذلك زيادة عدد درجات الحرية المتحكم بها من 1 إلى 10 وذلك غالبا بسبب تبديل الخوارزمية. وكذلك مقارنة مع دراسة (Shahid et al, 2020)، تمكن البحث الحالي في حالة حساسات اللمس من تحقيق سرعة أكبر في الحصول على نتائج مماثلة لدراسة shahid، وقد يكون ذلك بفضل حساسات اللمس الثنائية، التي تقلل من تعقيد العينة المقدمة للنموذج، إذ استخدم بحث Shahid 46 مكون لفضاء الحالة تمثل حالة الروبوت والكائن، في حين استخدم هذا البحث 20

معلمة يكون (10 منها قيم ثنائية)، ومن المعروف بالنسبة لخوارزميات RL يقلل هذا من الضجيج ويزيد من سرعة التدريب. أما مع حساسات الضغط احتاج البحث تقريبا لعدد خطوات متقارب كما في دراسة Shahid لتوليد نتائج مقاربة ولكن بالتحكم بعدد درجات حرية أكبر. بالنسبة لنتائج بحث (Kepes, 2017)، الذي حقق نسبة نجاح في الإمساك 100% مع يد بثلاث درجات حرية بعد تدريبها على مليون و 250 ألف خطوة تدريب ضمن المحاكاة بالإستفادة من حساسات اللمس وزوايا المفاصل وثم نقلها إلى يد خماسية الأصابع للتحكم في الإبهام والسبابة فقط وفشل النموذج عند محاولة نقله إلى أصابع أكثر، في حين حقق البحث الحالي نسبة النجاح ذاتها ولكن بشكل أسرع مع حساسات اللمس الثنائية 280 ألف خطوة وتمكن من التحكم بعدد درجات حرية أكبر موزعة على كامل أصابع اليد الخماسية. وقد يعود السبب في نوع الخوارزمية المختارة حيث تكون خوارزمية DDQN خوارزمية خارج السياسة وتعرف بأنها لا تتعامل جيدا مع البيئات المعقدة عالية الأبعاد، على عكس PPO المستخدمة في هذا البحث والتي تتعلم مباشرة داخل السياسة وتدعم الاستقرار والبيئات المعقدة. حقق نهج (Matak et al, 2022)، والذي استخدم صور العمق الملونة وحساسات اللمس للتحكم في التخطيط والتنفيذ لعمليات الإمساك بذراع ويد ذات 7 درجات حرية، نسبة نجاح 75% باستخدام منهجيات تحكم تقليدية في حين ساهم البحث الحالي بالحصول على نتائج أفضل ولدرجات حرية أكبر مع التعلم المعزز العميق. وهو ما يدعم النظريات التي تؤكد ضرورة إجراء أبحاث أكثر عن استخدام التعلم المعزز العميق في دراسات التحكم الروبوتي. احتاج بحث (Odeyemi et al, 2024)، 200 ألف خطوة لتحقيق عملية الإمساك بطرف ذو درجة حرية واحدة مع خوارزمية SAC وصور العمق الملونة كمدخلات، ضمن محاكي PyBullet. حيث حقق نجاح بنسبة 99%، ومع خوارزميات أخرى مثل DQN وPPO حقق 61% و 83% على التوالي. ويعزى معدل النجاح المرتفع الذي حققته خوارزمية SAC إلى مزاياها المتأصلة في التعامل مع مساحات العمل عالية الأبعاد. كما يمكن أن يُعزى معدل النجاح المنخفض الذي حققته خوارزمية DQN إلى حقيقة مفادها أن بنية التعلم Q المستخدمة بواسطة خوارزمية DQN معروفة بامتلاكها قضايا استقرار عند تطبيقها على مهام تحكم مستمر. فكما نلاحظ مع تقليل المدخلات وإلغاء المدخلات البصرسة وزيادة عدد درجات التحكم تمكن البحث الحالي من الحصول على نتائج أفضل، مع خوارزمية PPO، وتحسن النتائج من 83% في بحث Odeyemi إلى 100% في هذا البحث يؤكد النظرية المستخلصة والتي دعمها (Zhang et al, 2025)، بأن الصور تقلل نجاح الإمساك وبشكل خاص مع إضافة الضجيج ويدعم اللمس أمثلة النتائج وهو ما أكده أيضا (Imran et al, 2020) عند التدريب على الإمساك ضمن MuJoCo مع صور عمق ملونة حيث حصل على نسبة نجاح 78%، كما أكد ذلك مرة أخرى بحث (Zheng et al, 2023) عند تدريب روبوت مع 6 درجات حرية على إمساك جسم بوزن 500 غرام ضمن MATLAB باستخدام خوارزمية PPO مع مدخلات بصرية فقط حيث حصل على نسبة نجاح 90%، ومقارنة مع (100% أو 96%) التي حصل عليها في هذا البحث مع حساسات اللمس. كما تمكن البحث الحالي من حصول على تائج مقاربة لبحث (Puang et al, 2024)، الذي تمكن من الإمساك

بجسم متغير مركز الثقل مع طرف بدرجة حرية واحدة وحساسات اللمس كمدخلات، والذي حقق نسب نجاح أكبر من 95% مع خوارزمية SAC، وقلل متوسط فرق الضغط بالقيمة المطلقة إلى أقل من 0.2 نيوتن، فتمكن البحث الحالي من الحصول على نسب نتائج مقاربة مع عدد درجات حرية أكبر. بالنسبة لدراسات تعلم الآلة المعتمدة على IL، عند تدريب (Sharif et al, 2020) يد بخمسة أصابع وخمس درجات حرية مع SAC فشل الوكيل بالإمساك، في حين عند إضافة التعلم المحاكي تمكن من أداء المهمة، وبالتالي ساهم النهج الذي قدمه هذا البحث بأمثلة النتائج وربما يكون السبب وراء تحسن النتائج وإلغاء الحاجة للتعلم المحاكي وتحقيق النجاح لطرف بدرجات حرية أكبر، هو اعتماد البحث الحالي على المكافآت الكثيفة، في حين اعتمد Sharif على المكافآت المتفرقة والتي تعرف بأثرها الذي يخفض نسب النجاح. ويعرض الجدول (4-6)، مقارنة مع عدة دراسات.

الجدول 4-6 مقارنة النتائج الحالية مع الدراسات السابقة.

الدراسة	المهمة	المنهجية	نسبة النجاح
(Kepes, 2017)	إمساك ورفع	استخدم DDQN لتدريب قابض له 3 درجات حرية، ثم نقله ليد بسبع درجات حرية، على مهمة الإمساك ضمن محاكي MuJoCo، مع مدخلات قيم اللمس وزوايا المفاصل، عند محاولة التعامل مع أجسام متحركة أو درجات حرية أكبر لم يتمكن النموذج من النجاح.	**%100**
(Zhang et all, 2025)	الإمساك	استعمل قابض بأصبعين متوازيين له درجة حرية واحدة، مع ذراع للوصول، ضمن محاكي IssacGym للقيام بالإمساك والرفع مع مدخلات لمسية وبصرية، باستخدام خوارزميتي SAC وMPO مع حساسات لمس وقوة، وحققت الأخيرة نسب نجاح أفضل لذا نعرضها في عمود النتائج.	**%55 =MPO** **%66 = MPO**
(Matak et all, 2022)	التخطيط والتنفيذ للإمساك	استخدمت منهجيات متعددة لا تعتمد على RL لتعليم يد بسبع درجات حرية مع مدخلات صور العمق وحساسات اللمس.	**%75**
(Odeyemi et all, 2024)	الإمساك	استخدمت SAC وDQN وPPO لتدريب قابض بدرجة حرية واحدة ضمن محاكي PyBullet مع مدخلات صور العمق الملونة.	**%99 = SAC** **%61 =DQN** **%83 = PPO**
(Zheng, 2023)	الإمساك	استخدمت PPO ضمن Matlab لتدريب يد DOFBOT لها 6 درجات حرية على إمساك جسم وزنه 500 غرام على مهمة الإمساك.	**%90**
(Imran et all, 2020)	الإمساك	استخدمت DDPG لتدريب يد بخمس أصابع على الإمساك من خلال فتح وإغلاق الأصابع، ضمن MuJoCo مع صور العمق كمدخلات.	**%78**
(Ayukawa et all, 2021)	الإمساك	استخدمت PPO ضمن Unity مع ذراع بستة درجات حرية، مع مدخلات صور العمق، أظهرت النتائج عند 2 مليون خطوة إتمام المهمة بنجاح.	–
(Shahid et all, 2020)	الإمساك والوصول	استخدمت PPO ضمن MuJoCo لتدريب ذراع بسبع درجات حرية مع قابض بدرجة واحدة على مهمة الإمساك.	**%100**
(Timmers, 2018)	الإمساك والوصول	استخدمت CACLA لتدريب ذراع مع 6 درجات حرية وقابض بدرجة واحدة على مهمة الإمساك والصول، ضمن V-REP بدأ من الصفر، استمر 10 ساعات تدريب.	**%90**
(Puang et all, 2024)	الإمساك والرفع	يطُلب إمساك ورفع حاوية مستطيلة الشكل فيها حمل متحرك على طول الحاوية باستخدام محاكي DiffRedMax مع خوارزمية SAC. وضمن الواقع، مع قابض بدرجة حرية واحدة وأصبعين متوازيين.	**%95 مع 0.2 متوسط القوة الزائدة**

(Rivera et all, 2021)	توليد قبضات طبيعية	درب ذراع ADRIOT ذات 22 درجة حرية باستخدام خوارزمية NPG ضمن محاكي MuJoCo مع استخدام مساحة التآزر كتمثيل منخفض الأبعاد لحركات اليد، في كل خطوة يتم التحكم في زاوية المفاصل وموضع اليد وتقرأ السياسة هذه القيم مع المعلومات اللمسية. بهدف إنشاء قبضات طبيعية.	**88.38 %**
(Zeng et all, 2018)	الدفع والإمساك	درب روبوت UR5 مع قابض ثنائي الأصابع في بيئة محاكاة V-REP وفي الواقع مع مشاهدات بصرية، باستخدام خوارزمية Q – Learning.	**65.4%**
(Joshi et all, 2020)	الإمساك	استخدم المشاهدات البصرية للتحكم في روبوت Baxter بحيث يتمكن من الإمساك بمجسمات بسيطة ضمن بيئة Gazebo وعلى الروبوت الحقيقي، باستخدام خوارزمية DDQL مع Grasp Q Network مقترح ضمن الدراسة.	**87.%**
(Hundt et all, 2020)	الدفع والإمساك	اقترحت الدراسة استخدام إطار عمل جدولة المهام الإيجابية (Schedule for Positive Task – SPOT) مع خوارزمية Q-Learning حيث استخدمت المشاهدات البصرية للتحكم في قابض ذو أصبعين لأداء مجموعة متنوعة من المهام، ضمن بيئتي محاكاة وحقيقية.	**76%**
(Mohammed, 2020)	الإمساك والتموضع	درب روبوت UR5 مع قابض ثنائي الأصابع في بيئة محاكاة V-REP مع مشاهدات بصرية، باستخدام خوارزمية Q -Learning.	**74.3%**
(Al-Shanoon et all, 2021)	الإمساك	استخدام المشاهدات البصرية للإمساك بأجسام بسيطة بدأ من الصفر لروبوت UR5 ذو 6DOF، مع مقبض بأصبعين باستخدام (Deep Reinforcement Grasp Policy – DRGP) ضمن بيئة محاكاة V-REP وفي تجارب على الروبوت الحقيقي.	**62%**
(Peng et all, 2022)	الدفع والإمساك	استخدمت DQN مع كاميرا عمق ضمن V-REP لتدريب UR5 وقابض ثنائي.	**83.5%**
(Wu et all, 2023)	الدفع والإمساك	درب UR5 مع قابض ثنائي الأصابع ضمن V-REP و UR3 مع robotic-85 في الواقع مع مشاهدات بصرية، باستخدام DQN مما يساعد في أمثلة اتخاذ القرار.	**75.3%**
(Wang et all, 2024)	التكديس والإمساك	استخدم Q -Learning لتعلم ثلاث مهام وهي الإمساك، والتكديس، والدفع، لروبوت UR5 مع قابض ثنائي الأصابع ضمن V-REP مع مشاهدات بصرية.	**86%**
(Mohammed, 2021)	الدفع والإمساك	لتحقيق المهمة في سيناريوهات الفوضى والمرتبة والانسداد. قسم النهج إلى: (1) استخدام كاميرات متعددة لإعداد وجهات نظر متعددة لمعالجة مشكلة الانسداد؛ و(2) استخدام مشاهدات التغيير البصري على أساس اختلاف عمق البكسل. ضمن V-REP باستخدام UR5 مزود بمقبض فك متوازي.	**83.6%الازدحام 86.3%المنظمة 100%الانسداد**
(Franceschetti et all, 2020)	الإمساك والرفع	استخدم UR5 مع قابض بثلاث أصابع. (استخدمت ذات الروبوتات في المحاكاة والواقع). ضمن محاكي MuJoCo مع TRPO وDQNNAF وDDPG وVPG. لإمساك أسطوانة طولها 2 سم وكتلتها 0.5 كغ ثم وضعها على ارتفاع 30 سم فوق الجسم. في كل حلقة يتم أخذ موضع المفاصل كتغذية راجعة. أظهرت VGP وDDPG وDQNNAF فشل كبير بينما أظهرت TRPO إمساك ممتاز. لاحقا لذلك نجح نقل السياسة المقترحة في المحاكاة إلى تجارب الواقع دون الحاجة لإجراء أي تغيرات.	**100%**
البحث الحالي	الإمساك	درب طرف Handy الاصطناعي مع 10 درجات تحكم، خماسي الأصابع، لتولد قبضة أسطوانية واستخدم تعلم المناهج لتوليد قبضات كروية، ثم لتوليد قبضات أكثر دقة مع ثلاث أصابع، تم قراءة قيم المدخلات اللمسية (ثنائية، قوة)، مع زوايا المفاصل للتحكم في زاوية المفاصل، مع خوارزمية PPO ضمن محاكي Webots.	**اللمس الثنائي: 100% القوة: 96%**

6.6. الخاتمة:

تؤكد النتائج التي تم الحصول عليها على نجاح فرضية البحث القائلة أن التعلم المعزز من الممكن أن يكون نهجا مهما لابتكار أساليب تحكم متقدمة في طرف اصطناعي بالاعتماد على ردود الفعل اللمسية، وأكدت النتائج على الإمكانات الكبيرة لطرق التحكم باستخدام التعلم المعزز العميق، وأهمية المدخلات اللمسية في تحسين نسب النجاح في التحكم والإمساك لتوليد قبضات مثلى، بما يضمن تجاوز مشاكل المدخلات البصرية أو حتى الحاجة للعروض التوضيحية.

7. الفصل السابع: الاستنتاجات والآفاق المستقبلية.

7.1. مقدمة:

إن فقدان الأطراف يُعد من أكثر الأحداث المؤثرة على الصعيدين الجسدي والنفسي، إذ يغير بشكل جذري من نمط حياة الفرد وقدرته على التفاعل مع محيطه. وبفضل التطور المتسارع في مجالات التكنولوجيا والذكاء الاصطناعي، أصبحت هناك إمكانيات واعدة لاستعادة جزء كبير من هذه القدرات من خلال تصميم وتطوير أطراف اصطناعية ذكية قادرة على محاكاة الحركات البشرية الدقيقة والتفاعل مع البيئة. وقد لعب الذكاء الاصطناعي، وتحديدا مجال التعلم المعزز، دورا محوريا في هذا التقدم، حيث مكّن من تدريب وكلاء أذكياء قادرين على اتخاذ قرارات استراتيجية اعتمادا على التفاعل المستمر مع البيئة، بما يعزز من فرص النجاح في المهام الموكلة إليهم.

يُعتبر التعلم المعزز العميق نقلة نوعية في هذا المجال، لكونه يجمع بين قوة التعلم العميق في استخراج التمثيلات من البيانات، وقدرة التعلم المعزز على اكتساب السياسات المثلى من خلال التجربة والخطأ. وفي إطار هذا البحث، تم اقتراح نهج تدريبي مبتكر يمكّن وكيلا ذكيا من تعلم مهام الإمساك من خلال يد عالية درجات الحرية بالاستفادة من التغذية العكسية اللمسية وبدأ من الصفر، دون الحاجة إلى بيانات سابقة أو إشراف بشري مباشر. وقد بينت التجارب فعالية هذا النهج، وقد أظهرت السياسة الناتجة أداءً واعدا، إذ أظهرت السياسة المتولدة معدلات نجاح مرتفعة في مختلف السيناريوهات التي تم اختبارها، مما يؤكد الإمكانات العالية لهذا الأسلوب في تطوير نظم تحكم فعالة للأطراف الاصطناعية.

ورغم هذه النتائج المشجعة، ونتائج الدراسات السابقة التي تم استعراضها، إلا أن التحكم المتقدم في الروبوتات باستخدام التعلم المعزز العميق لا يزال يواجه تحديات كبيرة، خاصة فيما يتعلق بتعلم مهارات دقيقة، قوية، ومرنة بما يكفي لتلائم بيئات العالم الحقيقي. ومع ذلك، فقد أظهرت هذه الدراسة قدرة كبيرة لهذا المجال على التقدم نحو حلول أكثر واقعية وقابلية للتطبيق في مجال التحكم بالأطراف الاصطناعية.

7.2. الاستنتاجات النهائية:

يمكن تلخيص الاستنتاجات الرئيسية لهذا البحث في النقاط التالية:

1. **فعالية بيئة المحاكاة والتوزيع العشوائي للسيناريوهات:** أتاح استخدام بيئة محاكاة متنوعة وتوزيع عشوائي للظروف التدريبية (حصول التغيرات الطارئة على البيئة) للنموذج فرصة التعرض لتجارب غنية ومتعددة، مما ساهم في تعزيز قدرته على التعميم، وتقوية أدائه في مواجهة تغيرات بيئية مفاجئة أو غير مألوفة، وبالتالي أمثلة دقته ومرونته.
2. **أثر تعلم المناهج (Curriculum Learning) في تسريع الأداء:** أظهرت نتائج البحث أن اعتماد استراتيجية تعلم المناهج أدى إلى تسريع عملية التدريب وأمثلة أداء النموذج عند مواجهة مهام جديدة لم

يُدرّب عليها مسبقا، مع تقليل الحاجة للبدء بالتدريب من الصفر لكل مهمة. هذه النتائج تتماشى مع ما أشار إليه (Kepes, 2017) حول فعالية هذا النهج في اكتساب مهارات يصعب تعلمها بالطرق التقليدية.

3. **قوة تصميم هيكل المكافأة المقترح (Reward Structure):** أظهرت الاختبارات أن بنية المكافأة المصممة بعناية لحالتي الاختبار كانت قادرة على توجيه النموذج لتعلم سياسات إمساك دقيقة وفعالة، حتى في حال التعامل مع أجسام غير مألوفة أو غير مدروسة مسبقا، ما يدل على مرونة هذا التصميم وإمكاناته التطبيقية الواسعة. ليكون أساسا واعدا لتدريب نماذج على الإمساك بأجسام متنوعة، بما في ذلك الأجسام غير المألوفة أو التي تستخدم لأول مرة في سيناريوهات تدريب مختلفة.
4. **الإمكانات الواعدة لردود الفعل اللمسية:** أثبتت التجارب أن استخدام إشارات لمس ثنائية (binary tactile signals) يوفر معلومات كافية لتوليد حركات إمساك دقيقة ومعقدة، دون الحاجة لبيانات مجمعة مسبقا، أو توجيه بشري، وألغت الحاجة للاعتماد المفرط على الرؤية، وهو ما يمثل خيارا واعدا لتطوير أنظمة التحكم في الأطراف الاصطناعية، خاصة في البيئات التي تكون فيها المهمة صعبة، أو يصعب جمع بياناتها، أو الرؤية محدودة أو معرضة للضوضاء. في حين أفادت إشارات القوة بتقديم سياسات تحكم أكثر وثوقية إلا أنها أدت لبطء عام في أداء النظام، لأنه كان حذرا من تطبيق الضغط المفرط.
5. **دور حساسات اللمس في تعويض ضعف الرؤية** لعبت حساسات اللمس المدروسة في حالتيها دورا حاسما في تعويض أوجه القصور المرتبطة بالمعلومات البصرية، حيث أظهرت الدراسة أن إدخال بيانات لمسية يسهم بشكل كبير في رفع نسبة نجاح النماذج، خاصة في بيئات واقعية تعاني من تشويش بصري أو ظروف تصوير غير مثالية.

7.3. آفاق مستقبلية وتوصيات:

بناء على ما توصلت إليه هذه الدراسة، يمكن اقتراح عدد من المحاور المستقبلية التي من شأنها دعم النهج وتعزيزه:

1. **توسيع بيئة المحاكاة لتشمل الأجسام القابلة للتشوه (deformable objects):** لم تتعامل منصة المحاكاة المستخدمة في هذه الدراسة مع الأجسام الهشة، إذ أن Webots مصمم للتعامل مع الأجسام الصلبة، وهو ما يمثل قصورا محدودا في التقييم. لذا، يُقترح مستقبلا اعتماد منصة مثل DefGraspSim، التي طورتها NVIDIA عام 2023، والتي تستند إلى تحليل العناصر المحدودة (FEM)، مما يسمح بتقييم أكثر دقة لاستراتيجيات الإمساك خاصة في السياقات الجراحية أو الصناعية الحساسة.
2. **التحقق العملي من السياسات المدربة في بيئة واقعية:** تُعد خطوة نشر السياسات المدربة في طرف اصطناعي فعلي مرحلة ضرورية لتقييم فعالية النهج المقترح. فالتجارب الواقعية وحدها قادرة على كشف الفجوات بين الأداء في المحاكاة وبين الواقع، كما تتيح تحسينات تطبيقية ضرورية لجعل هذه النماذج قابلة

للاستخدام العملي في الأطراف الاصطناعية الذكية. لذا فنأمل مستقبلا العمل على اختبار هذه السياسات باستخدام طرف اصطناعي فعلي، بما يسهل التحقق العملي من النتائج ويعزز فرص التطبيق الواقعي لهذا النهج في مجال الأطراف الاصطناعية الذكية.

3. **دمج منهجيات جديدة مثل التعلم المعزز الضبابي (Fuzzy Reinforcement Learning):** في البيئات التي تتسم بالغموض وعدم اليقين المتأصلة في سيناريوهات العالم الحقيقي، فإن الاعتماد على القيم الدقيقة قد لا يكون كافيا. لذا فإن دمج منطق ضبابي ضمن بنية التعلم المعزز يُعد اتجاها واعدا. حيث يسمح هذا النوع من النماذج بالتعامل بمرونة أكبر مع البيانات الناقصة أو الغامضة، ويعزز قدرة النموذج على اتخاذ قرارات تكيفية في بيئات معقدة. ولا يوفر هذا التكامل للغموض في عملية التعلم نهجا أكثر تكيفا ومرونة فحسب بل يتماشى أيضا بشكل جيد مع التحديات التي تفرضها مهام الإمساك الروبوتية المعقدة والديناميكية.

7.4. الخاتمة:

لقد حاول هذا البحث المساهمة في سد فجوة حيوية في مجال تطوير سياسات تحكم ذكية للأطراف الاصطناعية، من خلال تقديم إطار تدريبي متكامل يعتمد على التعلم المعزز العميق، ويستفيد من تقنيات مثل تعلم المناهج، وبنية المكافأة الديناميكية، والمعطيات اللمسية. النتائج التي تم التوصل إليها تفتح آفاقا جديدة في هذا المجال، وتؤكد على أهمية مواصلة البحث والتطوير في سياق نقل هذه النماذج من المختبر إلى العالم الحقيقي، حيث تكمن الحاجة الملحة إلى أطراف اصطناعية أكثر ذكاء وفاعلية ومرونة.

المراجع:

1. Al-Shanoon, A., Lang, H., Wang, Y., Zhang, Y., & Hong, W., (2021). Learn to grasp unknown objects in robotic manipulation. Intelligent Service Robotics. 14(4), 571 – 582. doi:10.1007/s11370-021-00380-9.
2. Altuntas, E., & Sekeroglu, A., (2008). Mechanical Behavior and Physical Properties of Chicken Egg as Affected by Different Egg Weights. Journal of Food Process Engineering 33(1) 115–127. doi: 10.1111/j.1745-4530.2008.00263.x.
3. Argall, B. D., Chernova, S., Veloso, M., & Browning, B., (2009). A survey of robot learning from demonstration. Robotics and Autonomous Systems 57(5). 469–483.
4. Ayukawa, T., Yamaguchi, N., Fukuda, O., & Okumura, H., (2021). Reinforcement Learning of Electric Prosthetic Hand with Depth Camera in Simulation Environment. Bulletin of Networking Computing Systems and Software. 10(1), 17 - 19.
5. Barrasa, J. O., Montero, A. F., Ferrari, B., & Pastor, J. Y., (2021). Characterisation and Modelling of PLA Filaments and Evolution with Time. Polymers. 13(17), 2899. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8434208/.
6. Bick, D., (2021). Towards Delivering a Coherent Self-Contained Explanation of Proximal Policy Optimization. (Master's These). University of Groningen, The Netherlands.
7. Billard, A., Calinon, S., Dillmann, R., & Schaal, S., (2008). Robot Programming by Demonstration. In Siciliano, B., Khatib, O., (Eds.), Handbook of Robotics. Springer: Secaucus, NJ, USA. 1371–1394.
8. Bohg, J., Asfour, T., & Kragic, D. (2013). Journal IEEE TRANSACTIONS ON ROBOTICS. 1552-3098. doi: 10.1109/TRO.2013.228901.
9. Chervinskii, F., Rybnikov, A., Bogunowicz, D., & Vendidandi, K., (2020). Sim2Real for Peg Hole Insertion with Eye-in-Hand Camera. arXiv:2005.14401v1 [cs.RO].
10. Cordella, F., Ciancio, Al., Sacchetti, R., Davalli, A., Cutti, AG., Guglielmelli, E., & Zollo, L., (2016). Literature Review on Needs of Upper Limb Prosthesis Users. Frontiers in Neuroscience. 10,209. doi: 10.3389/fnins.2016.00209.
11. Coskun, M., Yildirim, O., & Demir, Y., (2022). Robotic Grasping in Simulation Using Deep Reinforcement Learning. In Proceedings of the 7th International Conference on Computer Science and Engineering (UBMK). doi:10.1109/UBMK55850.2022.9919482.
12. Danielsen, J. R., (2021). Vision Based Robotic Grasping in Simulation Using Deep Reinforcement Learning. (Master's Thesis). UIT the arctic university of Norway.
13. Evdokia, K. (2016) AI and Prosthetics. (Master's Thesis). Polytechnic University of Catalonia.
14. Feix, T., Romero, J., Schmiedmayer, H. B., Dollar, A. M., & Kragic. D., (2016). The Grasp Taxonomy of Human Grasp Types. IEEE Transactions on Human-Machine Systems. 46(1), 66–77.
15. Franceschetti, A., Tosello, E., Castman, N., & Ghidoni, S., (2020). Robotic Arm Control and Task Training through Deep Reinforcement Learning. arXiv:2005.02632v1 [cs.RO].
16. Gallouedec, M. Q., (2020). Deep Reinforcement Learning for Soft Objects Grasping. (Graduate report). Centrale Lyon École.
17. Gomes, N. M., Martins, F. N., Lima, J., & Wörtche, H. (2022). Reinforcement Learning for Collaborative Robots Pick-and-Place Applications: A Case Study. 3(1), 223–241. https://doi.org/10.3390/automation3010011.
18. Hundt, A., Killeen, B., Greene, N., Wu, H., Kwon, H., Paxton, C., & Hager, G. D., (2020). Good Robot!: Efficient Reinforcement Learning for Multi Step Visual Tasks with Sim to Real Transfer. IEEE Robotics and Automation Letters. 5 (4), 6724–6731.
19. Ibarz, J., Tan, J., Finn, C., Kalakrishnan, M., Pastor, P., & Levine, S., (2021). How to train your robot with deep reinforcement learning: lessons we have learned. The International Journal of Robotics Research. 40(4-5) 698–721.
20. Ijspeert, J., & Michel, O., (2004). Aibo Simulation in Webots and Controller Transfer to Aibo Robot. School of Computer and Communication Sciences Swiss Federal Institute of Technology Lausanne.
21. Imran, A., Escobar, W., & Barez, F., (2020). Design of an Affordable Prosthetic Arm Equipped with Deep Learning Vision-Based Manipulation. San Jose State University.

22. James, A. (2021). Estimation of Grip Pattern Using Reinforcement Learning for Prosthetic Arm. (Master's Thesis). Macquarie University.
23. Joshi, S., Kumra, S., & Sahin, F. (2020). Robotic grasping using deep reinforcement learning. In Proceedings of the IEEE 16th International Conference on Automation Science and Engineering (CASE) (pp. 1461–1466). Hong Kong, China. https://doi.org/10.1109/CASE48305.2020.9216953
24. Joubrane, K., Mnayer, D., Hamieh, T., Barbour, G., Talhouk, R. & Awad, E. (2019). Evaluation of Quality Parameters of White and Brown Eggs in Lebanon. American Journal of Analytical Chemistry, 10, 488-503. https://doi.org/10.4236/ajac.2019.1010035.
25. Kepes, M. (2017). Teaching Multi-fingered Robotic Hands to Grasp Simple Objects using Tactile Feedback-driven Deep Reinforcement Learning. (MENG individual project). Imperial college London.
26. Kilinc, O., & Montana, G., (2021). Reinforcement Learning for Robotic Manipulation Using Simulated Locomotion Demonstrations. Springer. Machine Learning. doi:10.1007/s10994-021-06116-1.
27. Kirtas, M., Tsampazis, K., Passalis, N., & Tefas, A. (2020). Deepbots: A Webots-based deep reinforcement learning framework for robotics. In Artificial Intelligence Applications and Innovations (AIAI 2020, IFIP AICT 584, pp. 64–75). Springer Nature Switzerland AG. https://doi.org/10.1007/978-3-030-49186-4_6.
28. Korber, M., Lange, J., Rediske, S., Steinmann, S., & Gluck, R., (2021). Comparing Popular Simulation Environments in the Scope of Robotics and Reinforcement Learning. arXiv:2103.04616v1 [cs.RO].
29. Korkmaz, S. (2018). Training a Robotic Hand to Grasp Using Reinforcement Learning. Arçelik A.Ş. ResearchGate. Access Date: 10/4/2022. url: 24 https://www.researchgate.net/publication/329389493.
30. Kormushev, P., Calinon, S., & Caldwell, D.G., (2013). Reinforcement Learning in Robotics: Applications and Real-World Challenges. Robotics. 2, 122 - 148. doi:10.3390/robotics2030122.
31. Li, Y., Wang, P., Li, R., Tao, M., Liu, Z., & Qiao, H., (2022). A Survey of Multi fingered Robotic Manipulation: Biological Results, Structural Evolvements, and Learning Methods. Frontiers in Neurorobotics. 16. Article 843267. doi:10.3389/fnbot.2022.843267.
32. Liu, R., Nageotte, F., Zanne, P., Mathelin, M., & Dresp-Langley, B., (2021). Deep Reinforcement Learning for the Control of Robotic Manipulation: A Focussed Mini-Review. Robotics. 10(1), 22. doi:10.3390/robotics10010022.
33. Luo, W., Wu, Z., & Yang, C., (2023). Strongly Coupled Fluid Structure Interaction Analysis of Aquatic Flapping Wings Based on Flexible Multibody Dynamics and the Modified Unsteady Vortex Lattice Method. Ocean Engineering. 281, 114921. https://doi.org/10.1016/j.oceaneng.2023.114921.
34. Mahmud, M., Kaiser, M. S., Hussain, A., & Vassanelli, S. (2018). Applications of Deep Learning and Reinforcement Learning to Biological Data. IEEE Transactions on Neural Networks and Learning Systems. doi:10.1109/TNNLS.2018.2790388.
35. Matak, M., & Hermans, T. (2022). Planning Visual-Tactile Precision Grasps Via Complementary Use of Vision and Touch. arXiv: 2212.08604v1. [cs.RO].
36. Mellatagshi, S. N. (2021). Learning Control of Robotic Arm Using Deep Q-Neural Network. University of Windsor. Electronic Theses and Dissertations. 8568. https://scholar.uwindsor.ca/etd/8568.
37. Mellatshahi, S. N. (2020). Learning Control of Robotic Arm Using Deep Q Neural Network. (Master's Thesis). University of Windsor.
38. Michel, O. (2004). Cyberbotics Ltd - Webots™: Professional mobile robot simulation. International Journal of Advanced Robotic Systems, 1(1), 39–42. ISSN 1729-8806.
39. Mohammed, M. Q., & Kweek, L. C., (2020). Pick and Place Objects in a Cluttered Scene Using Deep Reinforcement Learning. International Journal of Mechanical & Mechatronics Engineering 20(4), 50-57.
40. Mohammed, M. Q., Kweek, L. C., Chua, S. C., Aljaloud, A. S., Al-Dhaqm, A., Al - Mekhlafi, Z. G., & Mohammed, B. A. (2021). Deep Reinforcement Learning-Based Robotic Grasping in Clutter and Occlusion. Sustainability. 13(24), 13686. doi:10.3390/su132413686.
41. My Readers. (n.d.). Introduction to Artificial Intelligence. Retrieved February 12, 2025, from http://www.myreaders.info/html/artificial_intelligence.html.
42. Nia, N., Nassab, A., & Kaplanoglu, E. (2022). Reinforcement Learning-Based Grasp Pattern Control of Upper Limb Prosthetics in an AI Platform. In Proceedings of the 3rd International Informatics and Software Engineering Conference (IISEC). doi:10.1109/IISEC56263.2022.9998307.

43. Nizamli, F. M. (2020). Quality of Life Among Syrian Patients with War-Related Lower Limb Amputation at the Military Hospital in Lattakia. International Journal of Nursing Sciences. 7, 297 - 302. doi: 10.1016/j.ijnss.2020.05.0012.
44. Odeyemi, J., Ogbeyemi, A., Wong, K., & Zhang, W. (2024). On Automated Object Grasping for Intelligent Prosthetic Hands Using Machine Learning. Bioengineering. 11(2), 108.024, 11, 108. doi:10.3390/bioengineering11020108.
45. Orsula, A. (2021). Deep Reinforcement Learning for Robotic Grasping from Octrees Learning Manipulation from Compact 3D Observations. (Master's Thesis). Aalborg University.
46. Peng, G., Liao, J., Guan, S., Yang, J., & Li, X., (2022). A Pushing Grasping Collaborative Method Based on Deep Q-network Algorithm in Dual View Points. Scientific reports, 12(1), 3927. doi:10.1038/s41598-022-07900-2.
47. Puang, Y., Li, Z., Chew, M. C., Luo, S., & Wu, Y., (2024). Learning Stable Robot Grasping with Transformer-based Tactile Control Policies. arXiv:2407.21172v1 [cs.RO].
48. Rajeswaran, A., Kumar, V., Gupta, A., Vezzani, G., Schulman, J., Todorov, E., & Levine, S. (2018). Learning Complex Dexterous Manipulation with Deep Reinforcement Learning and Demonstrations. arXiv:1709.10087v2 [cs.LG].
49. Redman, T. R. (2016). "The Design of a Robotic Hand with Multiple Actuators for Children. (PhD thesis). University of Southampton.
50. Rivera, P., Anazco, A. V., & Kim, T. (2021). Object Manipulation with an Anthropomorphic Robotic Hand via Deep Reinforcement Learning with a Synergy Space of Natural Hand Poses. Sensors. 21(16), 5301. doi:10.3390/ s21165301.
51. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O., & OpenAI (2017). Proximal Policy Optimization Algorithms. arXiv:1707.06347v2 [cs.LG].
52. Sekkat, H., Moutik, O., Ourabah, L., Elkari, B., Chaibi, Y., & Tchakoucht, T. (2023). Review of Reinforcement Learning for Robotic Grasping: Analysis and Recommendations. Statistics Optimization and Information Computing. doi:10.19139/soic-2310-5070-1797.
53. Shahid, A. A., Piga, D., Braghin, F., & Roveda, L. (2022). Continuous Control Actions Learning and Adaptation for Robotic Manipulation Through Reinforcement Learning. Springer. Autonomous Robots. 46:483–498. doi:10.1007/s10514-022-10034-z.
54. Shahid, A. A., Roveda, L., Piga, D., & Braghin, F. (2020). Learning Continuous Control Actions for Robotic Grasping with Reinforcement Learning. In Proceedings of the 2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Toronto, Canada. pp.4066-4072. doi:10.1109/SMC42975.2020.9282951.
55. Sharif, M., Erdogmus, D., Amato, C., & Padir, T. (2020). Towards End-to End Control of a Robot Prosthetic Hand via Reinforcement Learning. Robotics and Biomechatronics (BioRob). 641-647. doi:10.1109/BioRob49111.2020.9224380.
56. Sherstan, C. (2015). Towards Prosthetic Arms as Wearable Intelligent Robots. (Master's Thesis). University of Alberta.
57. Shianifar, J., Schukat, M., & Mason, K. (2024). Optimizing Deep Reinforcement Learning for Adaptive Robotic Arm Control. arXiv:2407.02503v1 [cs.RO].
58. Sinha, J.K., Aziz, A., Ghosh, S. (2019). Homunculus. In: J. Vonk & T. Shackelford (Eds.), Encyclopedia of Animal Cognition and Behavior. Springer, Cham. https://doi.org/10.1007/978-3-319-47829-6_1427-1
59. Song, E. J., Lee, J. S., Moon, H., Choi, H. R., & Koo, J. C. (2021). A Multi-Curvature, Variable Stiffness Soft Gripper for Enhanced Grasping Operations. Actuators. 10(12), 316. doi:10.3390/act10120316.
60. Sun, T., Gong, L., Li, X., Xie, S., Chen, Z., Hu, Q., & Filliat, D. (2021). RobotDrlSim: A Real Time Robot Simulation Platform for Reinforcement Learning and Human Interactive Demonstration Learning. Journal of Physics. 1746, 012035. doi:10.1088/17426596/1746/1/012035.
61. Sutton, R. S., & Barto, A. G. (2018). Reinforcement Learning: An Introduction second edition. The MIT Press Cambridge, Massachusetts London, England.
62. Tao, S., Hu, Y., & Huang, H. (2023). Application of reinforcement learning algorithm on robot motion control and navigation in Webots simulator. J. Phys.: Conf. Ser. 2580 012036s. doi: 10.1088/1742-6596/2580/1/012036

63. Tatar, Y., Kablan, N., & Yurtseven, M. (2021) Incidence and Demographic Characteristics of Syrian Civil War-related amputations: A multi-centre study. Turk Journal of Physical Medicine and Rehabilitation. 67(1), 48-55. doi:10.5606/tftrd.2021.5058.
64. Thompson, C. R., Talla, R. R., Gummadi, J. C. S., & Kamisetty, A. (2019). Reinforcement Learning Techniques for Autonomous Robotics. Asian Journal of Applied Science and Engineering, 8, 1. 85 - 96.
65. Timmers, R. (2018). Learning to Grasp Objects with Reinforcement Learning. (Master's Thesis). University of Groningen.
66. Trararak, C., Hoang, T. T., & Cong-Kha, P. (2025). Pedestrian Avoidance Simulation by Deep Reinforcement Learning Using Webots. International Conference on Artificial Intelligence in Information and Communication (ICAIIC), Fukuoka, Japan, 2025, pp. 0734-0739, doi:10.1109/ICAIIC64266.2025.10920794.
67. Veiga, F., Akrour, R., & Peters, J. (2020). Hierarchical Tactile-Based Control Decomposition of Dexterous In-Hand Manipulation Tasks. Frontiers in Robotics and AI. 7. doi:10.3389/frobt.2020.521448.
68. Wang, F., Liu, Y., & Shi, M. (2024). Efficient Stacking and Grasping in Unstructured Environments. Journal of Intelligent and Robotic Systems. 110(57). doi:10.1007/s10846-024-02078-3.
69. Wu, L., Chen, Y., Li, Z., & Liu, Z. (2023). Efficient Push Grasping for Multiple Target objects in Clutter Environments. Frontiers in Neurorobotics. 17, 44 1188468. doi:10.3389/fnbot.2023.1188468.
70. Yu, C., Wang, P. (2022). Dexterous Manipulation for Multi-Fingered Robotic Hands with Reinforcement Learning: A Review. Front. Neurorobot. 16:861825. doi:10.3389/fnbot.2022.861825.
71. Zecca, M., Micera, S., Carrozza, M. C., Dario, P. (2002). Control of Multifunctional Prosthetic Hands by Processing the Electromyographic Signal. Critical Reviews in Biomedical Engineering, 30(4-6):27.
72. Zeng, A., Song, S., Welker, S., Lee, J., Rodriguez, A., & Funkhouser, T. (2018). Learning Synergies between Pushing and Grasping with Self-Supervised Deep Reinforcement Learning. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Madrid, Spain. 32, 4238–4245.
73. Zhang, B., Andrussow, I., Zell, A., & Martius, G. (2025). The Role of Tactile Sensing for Learning Reach and Grasp. University of Tübingen, Germany. arXiv:2502.20367v1.
74. Zhang, J., Wang, F., Yang, S., Zhang, Y., Song, H., & Li, M. (2015). Analysis of Drop Impact Behaviour of Eggs with Different Drop Height and Cushion Materials. Atlantis Press. AASRI International Conference on Industrial Electronics and Applications (IEA 2015).
75. Zhang, J., Zhang, W., Song, R., Ma, L., & Li, Y. (2020). Grasp for Stacking Via Deep Reinforcement Learning. 2020 IEEE International Conference on Robotics and Automation (ICRA), 2543-2549.
76. Zheng, Q., Peng, Z., Zhu, P., Zhao, Y., Zhai, R., & Ma, W. (2023). An Object Recognition Grasping Approach Using Proximal Policy Optimization with YOLOv5. IEEEACCESS. 11(9), 87330 - 87343. doi:10.1109/ACCESS.2023.3305339
77. Ziegler K, MacKenzie EJ, & Ephraim PL. Estimating the prevalence of limb loss in the United States: 2005 to 2050. Arch Phys Med Rehabil 2008;89(3): 422e9. doi:10.1016/j.apmr.2007.11.005

Abstract

Extensive research in the field of upper-limb prosthetics has led to the development of advanced control strategies. Nevertheless, designing prosthetic limbs capable of achieving high levels of dexterity and control comparable to that of the human hand remains a significant challenge. Recently, this field has witnessed a scientific breakthrough, driven by the technological integration of artificial intelligence (AI) and prosthetic sciences, resulting in the emergence of intelligent prosthetic limbs.

This thesis explores the application of deep reinforcement learning (a subfield of AI) in conjunction with curriculum learning to enable a high degree-of-freedom robotic hand to autonomously learn optimal motor behaviours through continuous interaction with its environment. This approach aims to enhance autonomy and motor intelligence, bringing the performance of the prosthetic hand closer to that of a natural human hand.

The significance of this research lies in the development of interactive prosthetic limbs capable of operating in diverse dynamic environments encountered by users, enabling efficient grasping actions without relying on traditional programming, and based solely on tactile information.

The experiments were conducted in the Webots simulation environment, utilizing the Proximal Policy Optimization (PPO) algorithm. The model used was the HANDY five-fingered prosthetic hand with 11 degrees of freedom, 10 of which were actively controlled during the experiments. During training, observations representing the real-time state of the hand were recorded. These included tactile feedback (studied in two configurations: binary contact sensing and force sensing, which measures pressure at contact points) as well as joint angles.

Based on these observations, the intelligent agent for each trained policy selected one of two actions: either incrementing the joint angle by a fixed value C, or maintaining the current position. A custom-designed reward function then reinforced the model based on the selected action, contributing to the gradual improvement of the policy's performance throughout the training episodes. The training process was evaluated using a set of quantitative and qualitative metrics.

To assess the effectiveness of the resulting decision-making policies, generalization experiments and robustness tests were conducted. Further improvements were achieved through curriculum learning, which led to a noticeable increase in success rates. Ultimately, the final decision policy achieved a 100% success rate in the binary tactile sensing scenario and 96% in the force sensing scenario.

These results demonstrate the high generalization capability of both the structure of the proposed approach and reward structures, highlighting their potential as a promising strategy for training prosthetic hands to interact effectively with a variety of objects.

Keywords: Upper Limb Prosthetics, Artificial Intelligence (AI), Deep Reinforcement Learning (DRL), Proximal Policy Optimization (PPO), Webots, Simulation.

Syrian Arab Republic

Damascus university

Faculty of Mechanical and Electrical Engineering

Department of Biomedical Engineering

Training a prosthetic hand model to grasp using deep reinforcement learning

A dissertation submitted in fulfilment of the requirement for the degree of Master in Biomedical Engineering

Prepared by

Eng. Thouraya Joubran Mouhssiena

Supervisors

Prof. Dr. Eng. Rasha Kazem Massoud

Dr. Eng. Faten Hafez Ajeeb

Academic year 2025