דוח מסכם למידת חיזוקים

מגישים :

נועה ענקי
רז אלבז
אושר דיגורקר

תוכן עניינים

פרק 1: מבוא

במערכות בריאות מודרניות, תזמון ניתוחים בחדרי ניתוח מהווה אתגר תפעולי מהותי. מדובר בבעיה מורכבת, בה יש לאזן בין מגבלות משאבים (כגון מספר חדרי ניתוח ושעות עבודה מוגבלות), לבין דרישות קליניות קריטיות – במיוחד כאשר חלק מהמטופלים מוגדרים כ"דחופים". החלטות שיבוץ שגויות עלולות להוביל לעיכובים, לשעות עבודה עודפות, ולפגיעה משמעותית בשביעות הרצון ובתוצאות הטיפול של המטופלים.

כדי להתמודד עם בעיה זו, נעשה שימוש בגישת למידת חיזוקים (*Reinforcement Learning – RL*), שיטה בה סוכן לומד מתוך אינטראקציה עם סביבה דינאמית כיצד לפעול בצורה מיטבית. בשונה מאלגוריתמים חמדניים או מבוססי חוקים, RL מאפשר למידה הדרגתית של מדיניות פעולה המביאה בחשבון השלכות עתידיות של כל החלטה.

בפרויקט זה נבנתה סביבה מותאמת המדמה חדרי ניתוח עם זרם משתנה של מטופלים, רמות דחיפות שונות, וזמני הגעה וניתוח מגוונים. הסוכן צריך ללמוד מתי וכיצד לשבץ כל מטופל כדי למקסם תגמול ארוך טווח – תגמול שמייצג את רמת היעילות והשירותיות של המערכת.

לאורך העבודה בחנו שלושה מודלים פופולריים של Deep RL:

DQN– אלגוריתם מבוסס Q-Value עם רשת נוירונים.

PPO– אלגוריתם אקטואלי מבוסס Policy Gradient עם שיפורים יציבותיים.

A2C– מודל מבוסס Actor-Critic עם עדכון סינכרוני.

בנוסף, השווינו את הביצועים מול שיטות פשוטות יותר – מדיניות אקראית וחמדנית – והצגנו ויזואליזציה מלאה של תהליך הלמידה, כולל סימולציה גרפית של הסוכן המנצח.

הפרויקט נשען בהשראתו על מאמר עדכני מאת Xu ואחרים (2023), שהציע שימוש ב־RL לצמצום תורים לאחר מגפה. במאמר זה נעשה שימוש ב־Deep Q-Learning על בסיס נתוני אמת, והמודל שנבחר הציג שיפור ניכר בשירותיות ובקיצור זמני המתנה. ברוח זו, פרויקט זה מדגים כיצד ניתן ליישם עקרונות דומים בסביבה מדומה ולבחון באופן מבוקר את ביצועי המודלים.

פרק 2: הגדרת הבעיה

הבעיה אותה ביקשנו לפתור עוסקת בניהול תזמון ניתוחים יומי בבית חולים, תחת מגבלות תפעוליות ובתנאים של אי-ודאות. בכל יום נדרש המערכת לשבץ כ־15 ניתוחים לאורך יום עבודה בן 480 דקות, ב־3 חדרי ניתוח זמינים. כל ניתוח מאופיין בשלושה פרמטרים: זמן הגעה (arrival time), משך ניתוח (duration), ורמת דחיפות בדידה בין 1 ל־3. רמות הדחיפות מוגרלות לפי הסתברויות קבועות מראש: 40% מהמטופלים מוגדרים כדחופים (דרגה 3), בעוד שהשאר נעים בין דחיפות בינונית לנמוכה.

הבעיה מתאפיינת בדינמיות גבוהה – המטופלים אינם זמינים מראש, אלא "נכנסים" לסביבה רק כאשר הגיע זמנם לפי זמן ההגעה שנקבע. תהליך זה מונע מהסוכן לדעת מראש את כל המידע, ומדמה תנאי חוסר וודאות הדומים לאלה של צוות תזמון אמיתי. בנוסף, קיימת מגבלה מהותית: אם משך ניתוח מסוים חורג מעבר לסיום יום העבודה, התוצאה היא קנס חמור, המדמה שעות נוספות והשלכות על תפעול המשמרת הבאה.

מרחב המצבים בנוי כ־dictionary הכולל חמישה מרכיבים: מצב חדרי הניתוח (כמות הזמן שנותרה לפעולה בכל חדר), השעה הנוכחית ביום, מספר המטופלים הממתינים, רשימת זמני ההמתנה של כל ממתין (עד 8 במקביל), ורשימת רמות הדחיפות שלהם. מרחב הפעולה הוא דיסקרטי וכולל את כל הצירופים האפשריים של (חדר, ממתין) – בנוסף לפעולה של המתנה.

פונקציית התגמול עוצבה בקפידה כדי לשקף את השיקולים הקליניים והתפעוליים. היא כוללת תגמול חיובי לשיבוץ מטופל (60 נקודות), תוספת תגמול של 40 נקודות עבור שיבוץ מקרה דחוף, וקנסות מצטברים על זמני המתנה ארוכים, בעיקר עבור מטופלים דחופים. ככל שהמתנה חורגת מ־15 דקות, הסוכן נענש בקנס גדול יותר, המדמה הידרדרות במצב המטופל. בנוסף, בכל דקה בה יש ממתין דחוף, נצבר קנס קטן פרופורציונלי לזמן ההמתנה. לבסוף, מטופלים שנותרו ללא שיבוץ בסוף היום גוררים קנס כבד, במיוחד אם מדובר בדחופים – כדי להבטיח שלכל פעולה תתלווה אחריות מערכתית כוללת.

הבעיה הנ"ל מציבה אתגרים אמיתיים לסוכן הלמידה: עליו לאזן בין תגמול מיידי לתכנון עתידי, לדעת מתי להמתין כדי לקבל החלטה טובה יותר, ומתי לשבץ כדי למנוע קנסות חמורים בהמשך. מדובר בבעיה קלאסית של למידת חיזוקים בסביבה סטוכסטית, שבה ההצלחה נמדדת לא רק על פי איכות ההחלטות הבודדות אלא גם על פי האסטרטגיה הכוללת שנבנית לאורך זמן.

פרק 3: תכנון הסביבה והמחקר

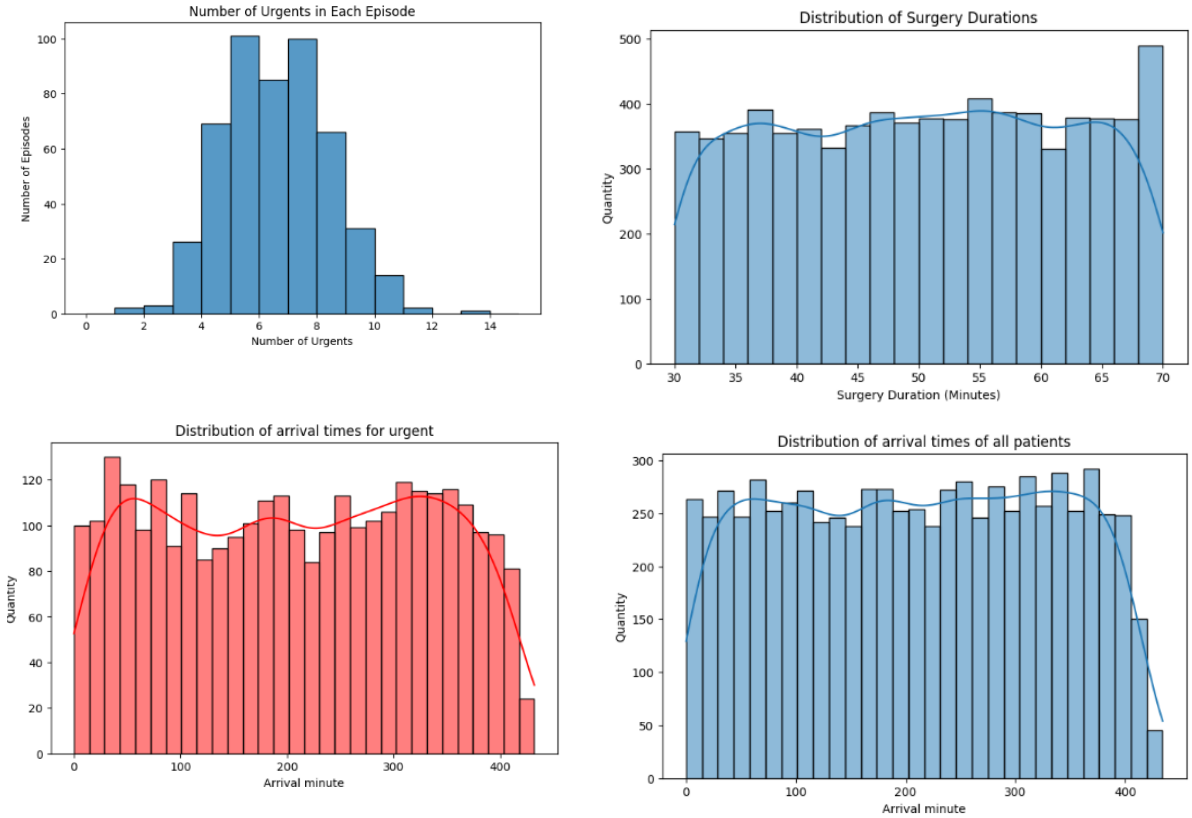
הבסיס לפיתוח הסוכן הלומד הוא סביבת הסימולציה אותה תכננו וכתבנו בקוד פתוח באמצעות ספריית gymnasium. הסביבה OperatingRoomEnv הוגדרה באופן מלא, תוך שליטה על כל רכיבי הדינמיקה: הגעת מטופלים, ניהול חדרים, פעולות חוקיות ולא חוקיות, ומעקב אחר ההיסטוריה של שיבוצים.

רשימות הניתוחים באפיזודה נוצרו על־ידי הפונקציה generate_surgeries, המגרילה 15 מטופלים לפי פרמטרים ריאליסטיים. עבור כל אחד מהם נקבע באופן אקראי משך ניתוח (30–70 דקות), זמן הגעה בטווח 0–480, ורמת דחיפות על פי חלוקה של 30% לדחיפות 1, 30% לדחיפות 2, ו־40% לדחיפות 3. הרשימה נחתכת וממויינת לפי זמן הגעה כדי להבטיח סדר כרונולוגי.

לצורך תכנון הניסויים והבנת הסביבה, הרצנו 500 אפיזודות שבהן נשמרו פרטי כל המטופלים. נאסף מידע על התפלגות זמני ההגעה, אורכי הניתוחים ורמות הדחיפות. תהליך זה סייע לנו להבין את הסביבה ולבחור מדדים מתאימים להשוואה בין המודלים.

בנקודה זו שילבנו גרפים חשובים המציגים את פיזור משתני הקלט:

Distribution of arrival times of all patients – מציג את פיזור זמני ההגעה, המעיד על פיזור אחיד לאורך היום. **Distribution of arrival times for urgent** – מתמקד במטופלים הדחופים בלבד, וממחיש כי אין להם ריכוז ברור בשעות מסוימות. **Distribution of Surgery Durations** – מציג את אורך הניתוחים, עם ממוצע סביב 45 דקות וסטיית תקן של 10 דקות. **Number of Urgents in Each Episode** – מתאר את פיזור מספר המטופלים הדחופים בכל אפיזודה, לרוב בטווח של 5–7.



הסביבה עצמה כוללת גם עטיפות (Wrapper) לשם הכנת הקלט למודלים: עטיפת `RandomPatientResetWrapper` מגרילה מחדש את המטופלים בתחילת כל אפיזודה, ו־`FlattenObservationWrapper` הופכת את התצפית לפורמט וקטורי התואם לרוב האלגוריתמים הקיימים.

תכנון הסביבה הושלם כאשר הוגדרו תנאים לעצירת האפיזודה (היום נגמר, כל המטופלים שובצו, וכל החדרים ריקים), והסוכן מוזמן ללמוד מתוך אינטראקציה שוטפת בסביבה זו כיצד לקבל החלטות אופטימליות. כך נבנה מערך מחקר מבוקר, הנשען על סימולציה חוזרת, גרפים סטטיסטיים, וניהול קפדני של קלטים ותגמולים.

פרק 4: שיטות

בשלב זה עברנו לבחינת שלוש שיטות למידת חיזוקים מתקדמות – DQN, PPO ו־A2C – שכל אחת מהן מייצגת גישה שונה לאינטראקציה עם הסביבה וללמידת מדיניות פעולה. המטרה המרכזית הייתה לבחון איזו שיטה מאפשרת לסוכן ללמוד מדיניות אפקטיבית ויציבה במציאות הדינמית של שיבוץ ניתוחים. כל מודל הוערך בשני שלבים – תחילה על בסיס הגדרות ברירת מחדל (baseline), ובהמשך לאחר חיפוש היפר־פרמטרים מדוקדק ואימון מורחב.

בשלב הראשון הרצנו כל אחד מהמודלים על פני 100 אפיזודות שונות, עבור ארבעה זרעים קבועים מראש (0, 42, 100, 999), כך שנוכל לבדוק את יציבותם תחת שונות מקרית. התוצאות הראשוניות הדגימו הבדלים ניכרים בין המודלים: בעוד ש־A2C הציג תגמולים חיוביים כבר בהתחלה, DQN ו־PPO הפיקו ערכים שליליים ובלתי יציבים במרבית האפיזודות. עם זאת, סטיית התקן הגבוהה ב־A2C העידה על חוסר יציבות ניכר, מה שדרש המשך בירור.

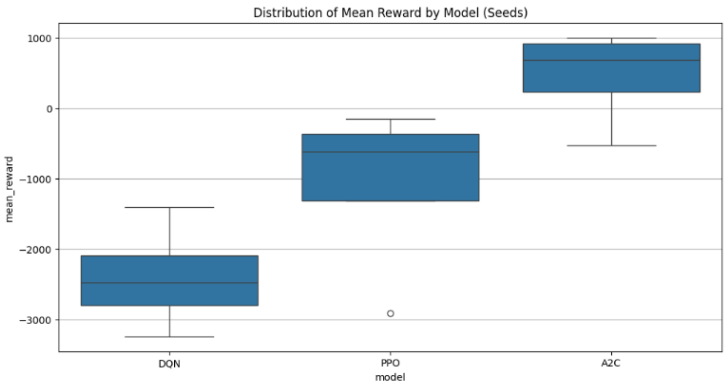
בהמשך, ביצענו חיפוש היפר־פרמטרים ייעודי לכל מודל. עבור DQN, A2C ו־PPO השתמשנו בשיטת Grid Search, במסגרתה נבדקו מאות שילובים של ערכים עבור קצב למידה, גודל batch, ערך gamma, ארכיטקטורת הרשת ועוד. המודל הטוב ביותר נבחר על בסיס ממוצע התגמולים לאחר 100 אפיזודות. כל מודל שאומן עם פרמטרים אופטימליים נבחן לאחר מכן על פני 500 אפיזודות אימון, ולאחריהן נמדד שוב בבדיקת ביצועים בת 100 אפיזודות רנדומליות לצורך השוואה סופית.

לעומת זאת, עבור A2C נדרשה גישה שונה. בשל חוסר היציבות והרגישות של המודל לפרמטרים, החלטנו לעבור לשיטת Bayesian Optimization תוך שימוש בספריית Optuna. שיטה זו איפשרה לנו למקד את החיפוש באזורים מבטיחים של מרחב ההיפר־פרמטרים, תוך שימוש בידע קודם שנצבר במהלך החיפושים. אחד הפרמטרים שהשפיעו בצורה דרמטית על הביצועים היה `n_steps`, המתאר את תדירות עדכון המדיניות – וכאשר הופחת ל־5 בלבד, נצפתה קפיצה דרמטית בתוצאות.

לצד המודלים הלומדים, שילבנו גם שני קווי בסיס (baselines) שאינם מתבססים על למידה, לצורך השוואה איכותית: מדיניות אקראית (Random Policy), שבה הסוכן בוחר פעולה באקראי מתוך מרחב הפעולה, ומדיניות חמדנית (Heuristic Policy) אשר שיבצה בכל רגע את המטופל הדחוף ביותר לחדר הראשון שהתפנה. הרצת 500 אפיזודות עבור כל אחת מהשיטות הללו הצביעה על הבדלים משמעותיים בין גישות חמדניות ללמידה עמוקה. המדיניות החמדנית הניבה תגמול שלילי בממוצע (-1759), ואילו המדיניות האקראית הציגה ביצועים חלשים מאוד אך עם שונות גבוהה.

לבסוף, יצרנו טבלה משווה הכוללת את כל שיטות הפעולה, עם התייחסות לממוצע התגמולים, סטיית התקן, זמן ההמתנה הממוצע, מספר ניתוחים דחופים שטופלו, אחוז הניתוחים בשעות נוספות

ועוד. טבלה זו מהווה בסיס לדיון שיופיע בפרק התוצאות.



Metric	DQN	PPO	A2C (Default)	A2C (Tuned)
Avg. total reward	**1095.61 ± 618.76**	850.63 ± 173.11	1047.58 ± 232.67	948.62 ± 192.55
Avg. waiting time (min)	**8.27 ± 5.13**	13.94 ± 6.44	11.79 ± 8.32	18.21 ± 6.68
Urgent surgeries served	7.13 ± 2.07	**7.41 ± 1.95**	7.11 ± 1.95	**6.79 ± 2.21**
Total surgeries per episode	17.94 ± 0.34	**18.00 ± 0.00**	**17.96 ± 0.20**	17.97 ± 0.30
Episodes with overtime (%)	**8.00%**	20.00%	12.00%	43.00%
Overtime surgeries per episode	**0.12**	0.29	0.17	0.65

פרק 5: תוצאות

בשלב זה של הפרויקט בוצעה השוואה בין ביצועי המודלים המאומנים על אפיזודות מבחן רנדומליות, שמטרתה לבחון את יכולת ההכללה של הסוכן ואת ביצועיו בסביבות חדשות ולא צפויות. כל מודל נבחן על פני 100 אפיזודות מבחן שונות, אשר הוגרלו באופן בלתי תלוי מאפיזודות האימון, במטרה לדמות שימוש אמיתי בסוכן לאחר הטמעה בעולם האמיתי – כלומר, בנסיבות שבהן הסוכן נתקל ברשימות ניתוחים חדשות, דפוסי הגעה שונים ורמות דחיפות אקראיות. תהליך זה מאפשר להעריך האם המודל למד מדיניות כללית שניתנת להכללה, או שמא הוא "זכר" את אפיזודות האימון בלבד. בכך, מדדנו לא רק את איכות הלמידה, אלא גם את העמידות של כל מודל בפני תנאים משתנים.

מודל DQN בלט במיוחד במדדי התוצאה. הוא השיג את התגמול הגבוה ביותר מכל המודלים, עם ממוצע של 1095.6 וסטיית תקן של 618.8, לצד זמן ההמתנה הנמוך ביותר שנמדד – 8.3 דקות בלבד. בנוסף, הוא הפגין אחוז נמוך מאוד של אפיזודות עם חריגות לשעות נוספות (8%) וכמות מזערית של ניתוחים שנעשו מעבר לשעה 480. עם זאת, סטיית התקן הגבוהה יחסית לתגמול שלו מרמזת על חוסר עקביות בביצועים, כאשר ישנם מקרים בודדים שבהם הסוכן נכשל באופטימיזציה.

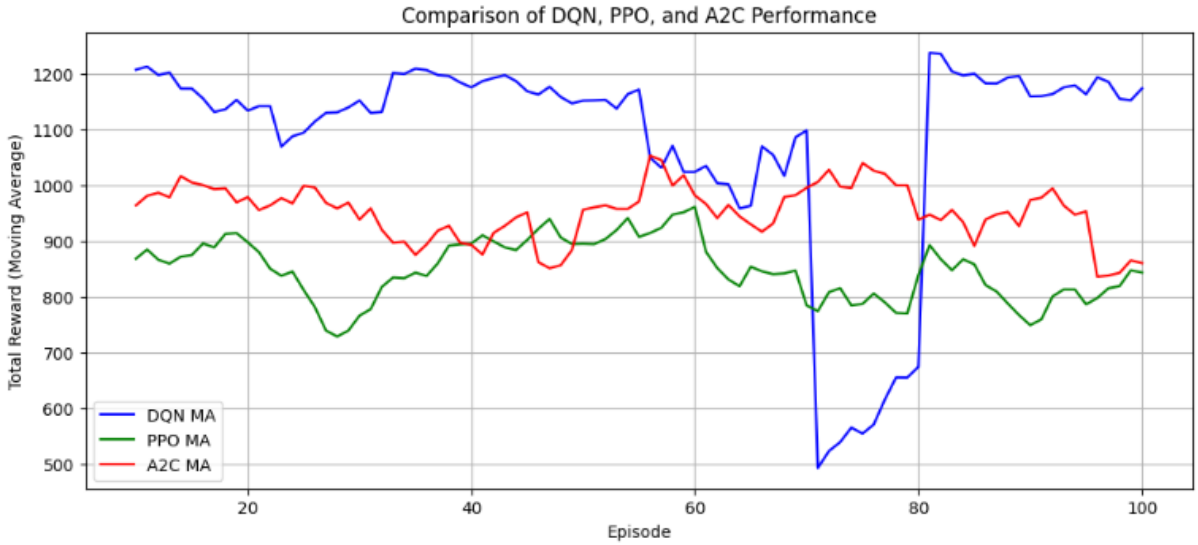
לעומתו, מודל A2C בגרסת ברירת המחדל הפגין התנהגות יציבה במיוחד: הוא הגיע לתגמול ממוצע של 1047.6 עם סטיית תקן נמוכה בהרבה (232.7), שמרה על רמת שירות טובה למטופלים הדחופים (7.11 בממוצע), והצליח לשמור על רמות סבירות של זמן המתנה (11.8 דקות) ואחוז שעות נוספות נמוך יחסית (12%). התנהגות זו מצביעה על כך שמודל זה למד מדיניות שמאזנת היטב בין יעילות תפעולית לבין גמישות והתמודדות עם עומסים.

גרסת A2C המכוונת (tuned) הובילה לירידה בתגמול הממוצע (948.6), יחד עם עלייה ניכרת בזמן ההמתנה (18.2 דקות) ובאחוז האפיזודות עם חריגת שעות (43%). כלומר, אף שבשלבי האימון נראתה מגמת שיפור, תוצאות המבחן חשפו כי הפרמטרים החדשים פגעו בביצועים הסופיים וגרמו למדיניות פחות אחראית תפעולית.

מודל PPO הפגין תגמול נמוך משמעותית מהאחרים (850.6) עם זמן המתנה ממוצע של 13.9 דקות. אף שהוא טיפל במספר גבוה יחסית של מטופלים דחופים, הוא עשה זאת במחיר של אחוז גבוה של חריגות לזמן נוסף (20%), מה שמעיד על נטייה לדחוף עוד ניתוחים גם בשלב מאוחר של היום – דבר שפוגע ביעילות הכוללת ובאיזון שבין איכות לשירות.

במבט כללי ניתן לראות שמודל DQN בחר בפתרון אגרסיבי אך אפקטיבי – תגמולים גבוהים מאוד בזכות שיבוץ יעיל ומהיר, לצד מינימום שעות נוספות. עם זאת, הוא סבל מתנודתיות בביצועיו, דבר שהופך אותו לפחות צפוי. לעומתו, מודל A2C (ברירת מחדל) אומנם לא הגיע לשיא התגמולים, אך הפגין עקביות מרשימה ויציבות שהופכות אותו לאופציה אמינה יותר בפריסה רחבה.

בגרף הבא ניתן לראות את ההשוואה בין שלושת המודלים לאורך 500 אפיזודות פר מודל.

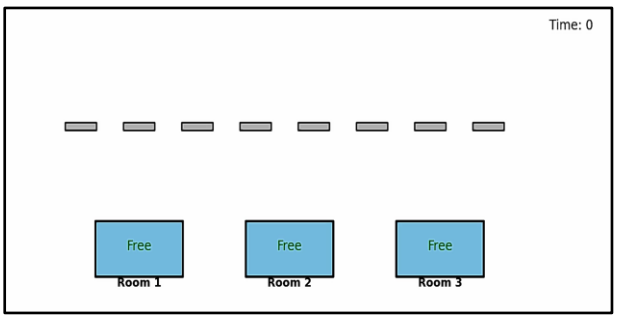


בהמשך הפרויקט, בחרנו להמחיש את ביצועיו של מודל A2C באמצעות הדמיה גרפית מלאה. הסימולציה תיעדה אפיזודה שלמה, צעד אחר צעד, תוך הצגת מצב חדרי הניתוח, זמני ההמתנה והחלטות השיבוץ של הסוכן. בתצוגה החזותית היה ניתן לראות כיצד הסוכן פועל תוך שמירה על סדר, מתן עדיפות למטופלים דחופים והימנעות משיבוצים המובילים לחריגות מיותרות. ההחלטות שהתקבלו היו מדודות, לעיתים אפילו כאלה שכללו המתנה יזומה במקום פעולה מיידית, וזאת במטרה להשיג הקצאה טובה יותר מספר צעדים קדימה.

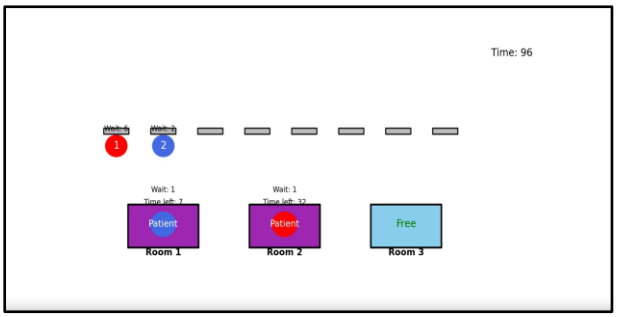
גרסה מונפשת של הסימולציה צורפה לסרטון הסבר שהוכן בפרויקט, ובדו"ח זה נכללו מספר צילומי מסך הממחישים את אופן פעולת הסוכן בזמן אמת. שילוב המדדים הכמותיים עם התבוננות איכותית בהתנהגות בזמן פעולה איפשרו לקבל תמונה שלמה על יכולות המודל.

לצורך המחשה ויזואלית של אופן הפעולה של הסוכן, צורפה סימולציה גרפית המדמה אפיזודה מלאה בזמן אמת. שלוש תמונות נבחרות מתוך הסימולציה ממחישות כיצד נראית הסביבה, מהם הנתונים שהסוכן רואה, וכיצד הוא מקבל את החלטות השיבוץ שלו לאורך הזמן.

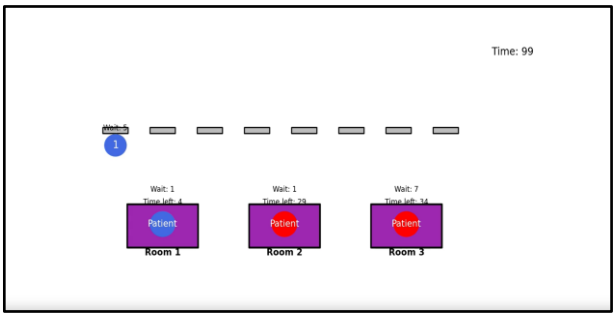
בצילום הראשון, ניתן לראות את הסביבה בתחילת האפיזודה – שלושה חדרי ניתוח פנויים, שמונה מושבים בתור ההמתנה, כאשר טרם הגיעו מטופלים. מצב זה מדגים את נקודת האתחול, בה הסוכן ממתין להגעת המטופלים ומתחיל לצבור תצפית על הסביבה.



בתמונה השנייה נראים שני מטופלים ממתינים – אחד דחוף בצבע אדום, ואחד לא דחוף בצבע כחול. שני חדרים כבר מאוכלסים, והשלישי פנוי. ניתן להבחין שהסוכן מקבל מידע על זמן ההמתנה של כל מטופל, וכי זמן ההמתנה של המטופל הדחוף ארוך יחסית. זהו שלב קריטי בהחלטה: האם להעדיף שיבוץ מידי לפי תור ההגעה, או להפעיל שיקול דחיפות.



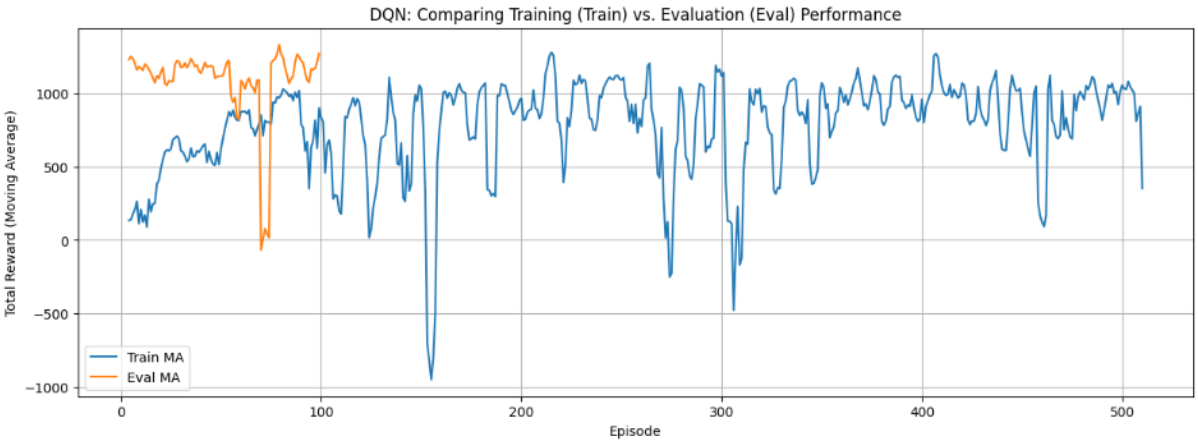
בתמונה השלישית מוצגת תוצאה של מדיניות שנלמדה היטב: כל שלושת חדרי הניתוח תפוסים, כאשר המטופל הדחוף שובץ בעדיפות גבוהה והמטופל הלא דחוף ממתין. בכך ניתן לראות בבירור כיצד הסוכן פועל לפי העדפת תגמול מצטבר – הוא מוכן לדחות שיבוץ מיידי של מטופל פחות דחוף, כדי להבטיח שהמשאבים ינוצלו קודם למקרים קריטיים יותר.



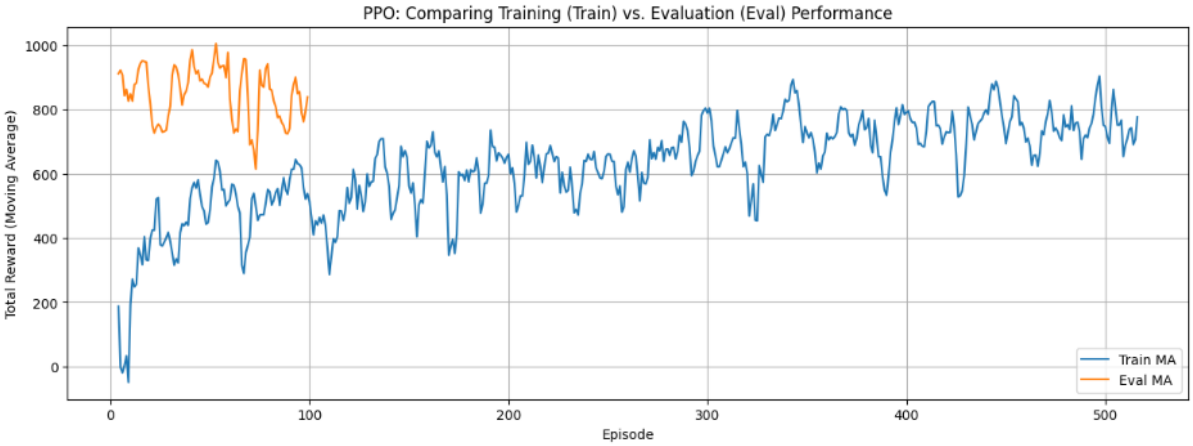
רצף הצילומים ממחיש שהסוכן לא פועל באופן חמדני או אקראי, אלא מפעיל מדיניות שנלמדה לאורך מאות אפיזודות, במסגרתה נשקלים דחיפות, זמני המתנה, מצב החדרים והשלכות של החלטות מוקדמות על המשך היום.

מעבר לניתוח הביצועים המספריים, ביצענו גם השוואה חזותית של מהלך הלמידה של כל סוכן. ההשוואה בוצעה באמצעות גרפים המתארים את התקדמות התגמול המצטבר במהלך האימון, לעומת הביצועים שנמדדו על אפיזודות מבחן שלא נראו קודם לכן. גרפים אלה שימשו לבחינת יציבות הלמידה, כושר ההכללה, והפערים בין הסתגלות לתנאים מוכרים לבין תפקוד מול מצבים חדשים.

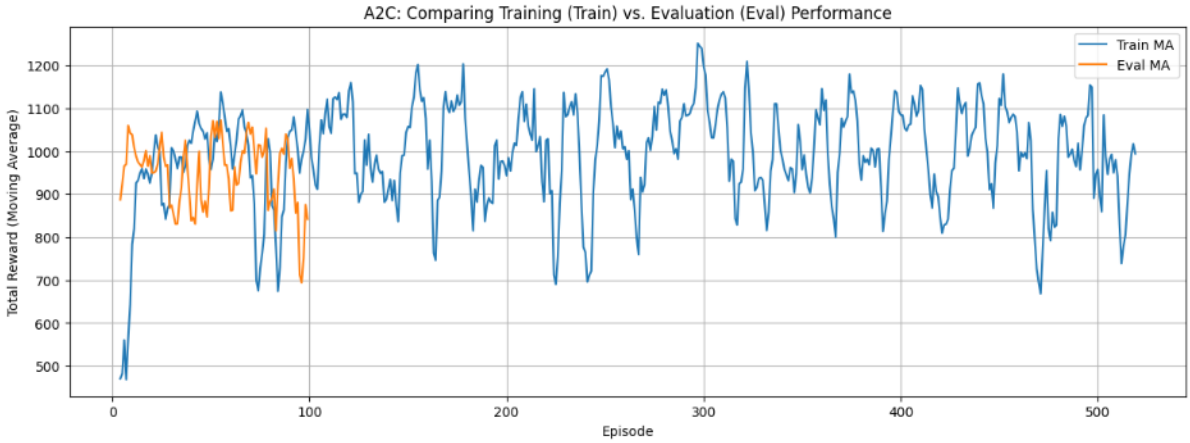
מודל DQN הפגין במהלך האימון תנודתיות ניכרת — עם עליות חדות וירידות לסירוגין, אך גם השגת תגמולים גבוהים במיוחד בשיאי הלמידה. בהערכת המבחן הוא הפגין יציבות גבוהה יחסית, מה שמרמז על כך שלמרות התנודתיות, המודל מצליח להכליל בצורה טובה לסביבות חדשות. ההתנהגות הזו תואמת את ממצאי פרק הביצועים, בהם DQN הצטיין בתגמול הגבוה ביותר.



מודל PPO, לעומתו, התקדם בהדרגה ובאופן מדוד במהלך שלב האימון. אמנם תגמוליו לא הגיעו לגבהים של DQN, אך עקומת הלמידה הייתה חלקה יותר. בשלבי המבחן, ניתן היה לראות ירידה קלה בתגמולים יחסית לאימון – מה שמעיד על כושר הכללה מתון, אך לא מבריק. המודל למד התנהגות כללית סבירה, אך נותר מאחור לעומת המתחרים.



מודל A2C בהגדרות ברירת המחדל הציג את ההתנהגות החלקה והיציבה ביותר מכלל המודלים. כבר בשלבים מוקדמים של האימון ניתן היה לראות התכנסות עקבית לתגמולים גבוהים, ללא קפיצות פתאומיות. גם בביצועי המבחן ניכר כי המודל שומר על רמת ביצועים כמעט זהה, מה שמעיד על מדיניות אחידה, גמישה, ועמידה לשונות מקרית.



ניתוח זה מחזק את המסקנות הקודמות: DQN מתאים לפרויקטים בהם הביצועים המרביים הם העדיפות הראשונה, גם אם המחיר הוא תנודתיות מסוימת. לעומתו, A2C מהווה מודל אמין ויציב, שמתאים יותר להטמעה בסביבות תפעוליות הדורשות עקביות, ריסון והכללה אמינה. PPO ממוקם בין השניים – כמודל מאוזן, אך פחות בולט באחד מההיבטים.

פרק 6: דיון וסיכום

הפרויקט הנוכחי ביקש להתמודד עם אתגר מורכב של תזמון ניתוחים תחת מגבלות תפעוליות ורמות דחיפות משתנות. לאורך שלבי העבודה בנינו סביבה מדויקת שמדמה מציאות קלינית משתנה, אימנו שלושה מודלים מבוססי למידת חיזוקים והשווינו את ביצועיהם, הן ברמה הכמותית והן בהתנהגות בזמן אמת. התוצאות הדגימו באופן ברור את הפערים בין הגישות השונות – הן ביעילות, הן ביציבות, והן ביכולת ההכללה.

המודל DQN בלט במיוחד ביכולתו למקסם תגמול בטווח הארוך. הוא הפגין את הביצועים הטובים ביותר במדדי התוצאה – תגמול כולל, זמני המתנה קצרים ומיעוט חריגות לשעות נוספות. עם זאת, תצפיות מהאימון חשפו כי המודל סובל מתנודתיות גבוהה, המתבטאת ברגישות למאפייני אפיזודות שונים או ערכי seed משתנים. המשמעות היא שהשגת הביצועים המרביים כרוכה במחיר של חוסר עקביות, דבר שעשוי להיות משמעותי במערכות בריאות בהן נדרשת אמינות מרבית.

מנגד, מודל A2C בגרסת ברירת המחדל הציג רמת יציבות גבוהה מאוד. הוא למד מדיניות אחידה ומאוזנת שהתמודדה היטב עם שונות בין אפיזודות, הן באימון והן במבחן. הוא אמנם לא הגיע לרמות התגמול של DQN, אך שמר על התנהגות אחראית, תפקוד מובהק מול דחופים, ותגובה מאוזנת לעומסים. במערכות רפואיות אמיתיות, בהן צפויה שונות יומית גבוהה, מודל מסוג זה עשוי להיות עדיף דווקא בשל תכונותיו היציבות.

מודל PPO הוכיח עצמו כמועמד ביניים – הוא הצליח להשתפר לאורך זמן, והציג התנהגות סבירה ויחסית מאוזנת, אך לא הצליח להגיע לרמת הדיוק או הגמישות של שני המודלים האחרים. נראה כי בפורמט הספציפי של הבעיה הוא נותר פחות מותאם.

ניתן לסכם את מסקנות הפרויקט כך:

המודלים של למידת חיזוקים הוכיחו יכולת גבוהה להבין את מבנה הבעיה, לפתח מדיניות פעולה חכמה, ולשפר באופן ממשי את איכות קבלת ההחלטות בהשוואה למדיניות אקראית או חמדנית. כל אחד מהם מביא יתרון שונה, בהתאם להקשר השימוש.

בהתאם לכך, ההמלצה הסופית לבחירת המודל תלויה בהקשר היישומי:

אם המטרה היא מקסימום יעילות תפעולית, גם במחיר של שונות בין אפיזודות –

אזי מודל **DQN** הוא הבחירה המתאימה.

לעומת זאת, אם הסביבה דורשת יציבות, חזרתיות והימנעות מחריגות,

מודל **A2C** (ברירת מחדל) הוא המתאים ביותר, גם במחיר ירידה קלה בתגמול.

שני המודלים מציגים ביצועים מעשיים ואמינים, ומהווים בסיס מוצק למחקר עתידי ואף ליישום במערכות תכנון קליניות בעולם האמיתי.

בהמשך לעבודה זו, ניתן לזהות מספר כיוונים להעמקה והרחבה:

ראשית, ניתן להרחיב את הסביבה כך שתכלול משתנים קליניים נוספים, כמו זמינות צוות רפואי, מגבלות בציוד, או הבדלים ברמות דחיפות בתוך קבוצת המטופלים הדחופים עצמם. שילוב של נתוני אמת או סימולציה מבוססת על בתי חולים קיימים עשוי לקרב את המודל למציאות תפעולית.

כיוון נוסף הוא פיתוח ממשק הפעלה אינטראקטיבי, שיאפשר למתכנן אנושי להתערב בהחלטות הסוכן ולבחון חלופות. גישה זו עשויה להוביל למערכות תומכות החלטה המשלבות בין בינה מלאכותית לשיקול דעת מקצועי.

לבסוף, בהיבט המחקרי, ניתן להרחיב את ניתוח היציבות והרגישות של הסוכנים לפרמטרים שונים, ובכך להעמיק את ההבנה של התנאים בהם כל מודל מצליח או כושל.

פרק 7 : ביבליוגרפיה

u, H., Fang, Y., Chou, C.-A., Fard, N., & Luo, L. (2023). A reinforcement learning-based optimal control approach for managing an elective surgery backlog after pandemic disruption. Health Care Management Science, 26, 430–446