



UNIVERZITET U NOVOM SADU
FAKULTET TEHNIČKIH NAUKA U
NOVOM SADU



**Automatski sistem upravljanja uređajima
pametne kuće za balansiranje troškova
električne energije i udobnosti stanovanja
zasnovan na metodama dubokog učenja**

DOKTORSKA DISERTACIJA

Mentori:

prof. dr Jelena Slivka

prof. dr. Jovana Vidaković

Kandidat:

Aleksandra Kaplar

Novi Sad, 2026. godine

КЉУЧНА ДОКУМЕНТАЦИЈСКА ИНФОРМАЦИЈА¹

Врста рада:	Докторска дисертација
Име и презиме аутора:	Александра Каплар
Ментор (титула, име, презиме, звање, институција):	др Јелена Сливка, редовни професор, Факултет техничких наука, Универзитет у Новом Саду др Јована Видаковић, ванредни професор, Природно-математички факултет, Универзитет у Новом Саду
Наслов рада:	Аутоматски систем управљања уређајима паметне куће за балансирање трошкова електричне енергије и удобности становања заснован на методама дубоког учења
Језик и писмо рада:	Српски (латиница)
Физички опис рада:	Унети број: Страница:200 Поглавља: 7 Референци:146 Табела:25 Слика: 73 Графика:0 Прилога:0
Научна област:	Електротехничко и рачунарско инжењерство
Ужа научна област (научна дисциплина):	Примењене рачунарске науке и информатика
Кључне речи / предметна одредница:	дубоко учење, системи за управљање енергијом у паметним кућама, двоструко дубоке Q-мреже, оптимизација проксималне политике, велики језички модели, промпт инжењеринг, страгеттије дефинисања упита без и са неколико примерима
Апстракт на језику рада:	Развијено је модуларно симулационо окружење које омогућава реалистичну симулацију рада кућних уређаја и једноставно дефинисање различитих конфигурација паметних кућа и екстерних услова у којима се оне налазе. Имплементирана су два приступа симулацији рада уређаја: симулација са константном потрошњом енергије током рада уређаја и симулација са варијабилном потрошњом енергије у оквиру радног циклуса, чиме се омогућава веродостојније представљање реалних система. Претходна истраживања су се претежно ослањала на симулације у

¹ Аутор докторске дисертације потписао је и приложио следеће Обрасце:

5б – Изјава о ауторству;

5в – Изјава о истоветности штампане и електронске верзије докторског рада и дозвола за објављивање личних података;

5г – Изјава о коришћењу.

Ове Изјаве се чувају у институцији у штампаном и електронском облику и не корице се са радом.

којима је потрошња уређаја моделована као константна током њиховог рада. Међутим, резултати ове дисертације указују да избор начина моделовања потрошње уређаја може имати значајан утицај на перформансе разматраних приступа за оптимизацију рада кућних уређаја.

Дефинисана је вишециљна функција награде која обједињује трошкове електричне енергије и комфор станара. Комфор станара је моделован у монетарним јединицама и представља меру њихове спремности да плате коришћење кућних уређаја у одређеним временским интервалима. На овај начин омогућено је директно обједињавање комфора и трошкова електричне енергије у оквиру јединствене функције награде, уз јасно економско тумачење оба критеријума. За разлику од постојећих приступа, у којима се функција награде најчешће дефинише као пондерисана сума трошкова и комфора са хеуристички одабраним тежинским коефицијентима, предложени приступ омогућава методолошки утемељено балансирање ова два циља.

Као решења за оптимизацију постављеног проблема разматрани су модели засновани на учењу поткрепљивањем, укључујући двоструку дубоку Q-мрежу и оптимизацију проксималне политике, као и приступи засновани на употреби претренираних великих језичких модела (BJM). У тренутку писања ове дисертације није пронађена ниједна студија која је разматрала примену BJM у контексту оптимизације времена рада кућних уређаја, нити њихову ефикасност у симулацији процеса доношења одлука аналогних онима који се користе у учењу поткрепљивањем.

Предложена решења упоређена су са референтним моделом заснованим на знању дефинисаном од стране људског експерта. Као критеријум успеха постављено је да разматрана решења остваре исте или боље перформансе у односу на референтни модел, уз доношење одлука које су оправдане из перспективе људског експерта. Резултати показују да модел заснован на оптимизацији проксималне политике испуњава постављене критеријуме и остварује најстабилније и најробусније перформансе у поређењу са осталим разматраним приступима, док двострука дубока Q-мрежа не испуњава постављене критеријуме. Приступи засновани на великим језичким моделима (BJM) нису у потпуности задовољили постављене критеријуме, али су остварили перформансе упоредиве са референтним моделом, што указује на њихов истраживачки потенцијал.

Предност приступа заснованих на BJM у односу на методе учења поткрепљивањем огледа се у елиминацији потребе за дуготрајним процесом тренирања. Такође, треба истаћи да су у оквиру ове дисертације разматране основне

	варијанте ових приступа, те да постоји простор за унапређење њихових перформанси. Резултати су показали да перформансе и стратегије разматраних приступа у великој мери зависе од годишњег доба. За разлику од ове дисертације, већина постојећих радова не разматра утицај сезонских разлика. Добијени резултати додатно указују на предност предложених приступа у односу на референтни модел, која се огледа у њиховој способности аутоматске адаптације на променљиве услове окружења, док би систем заснован на знању, дефинисан од стране људског експерта, захтевао редизајн при свакој значајнијој промени окружења.
Датум прихватања теме од стране надлежног већа:	20.11.2025.
Датум одбране: (Попуњава накнадно институција)	
Чланови комисије: (титула, име, презиме, звање, институција)	Председник: др Милан Видаковић, редовни професор, Факултет техничких наука, Нови Сад Члан: др Александар Ковачевић, редовни професор, Факултет техничких наука, Нови Сад Члан: др Никола Војновић, ванредни професор, Факултет техничких наука, Нови Сад Члан: др Милош Радовановић, редовни професор, Природно-математички факултет, Нови Сад Ментори: др Јелена Сливка, редовни професор, Факултет техничких наука, Нови Сад др Јована Видаковић, ванредни професор, Природно-математички факултет, Нови Сад
Напомена:	

KEY WORD DOCUMENTATION²

Document type:	Doctoral dissertation
Author:	Aleksandra Kaplar
Supervisor (title, first name, last name, position, institution)	PhD, Jelena Slivka, full professor, Faculty of Technical Sciences, University of Novi Sad PhD, Jovana Vidaković, associate professor, Faculty of Technical Sciences, University of Novi Sad
Thesis title in English:	An automated system for managing smart home devices to balance electricity costs and living comfort based on deep learning methods
Language and script:	Serbian (latin)
Physical description:	Number of: Pages:200 Chapters:7 References:146 Tables:25 Illustrations:73 Graphs:0 Appendices:0
Scientific field:	Electrical and computer engineering
Scientific subfield (scientific discipline):	Applied Computer Science and Informatics
Subject, Key words:	deep learning, house energy management systems, double deep Q-networks, proximal policy optimization, large language models, prompt engineering, zero-shot and few- shot query definition strategies
Abstract in English:	Smart homes are becoming increasingly prevalent in everyday life. The integration of renewable energy sources and energy storage systems provides significant potential for more economical and sustainable electricity consumption in smart homes. However, manual control of appliances in such systems requires time as well as a certain level of technical expertise, which suggests that automation of appliance management systems could substantially reduce the burden on residents. This dissertation addresses the problem of automatic appliance

² The author of the doctoral dissertation has signed the following Statements:

56 – Statement on the authorship,

5B – Statement that the printed and e-version of the doctoral dissertation are identical and authorization to use personal data,

5r – Copyright statement.

The paper and e-versions of Statements are held at the institution and are not included into the printed thesis.

	control in a smart home, with the objective of simultaneously optimizing electricity consumption and occupant comfort. A modular simulation environment has been developed, enabling realistic simulation of household appliances as well as straightforward definition of various smart home configurations and external environmental conditions. Two approaches to appliance consumption modeling are implemented: simulation with constant energy consumption during appliance operation and simulation with variable energy consumption within the operating cycle, allowing for a more faithful representation of real-world systems. Previous studies have predominantly relied on simulations in which appliance energy consumption is modeled as constant throughout operation. However, the results of this dissertation indicate that the choice of appliance consumption modeling approach can have a significant impact on the performance of the considered methods for appliance operation optimization. A multi-objective reward function is defined that jointly accounts for electricity costs and occupant comfort. Occupant comfort is modeled in monetary terms and represents a measure of residents' willingness to pay for the operation of household appliances during specific time intervals. This formulation enables the direct integration of comfort and electricity costs within a unified reward function, providing a clear economic interpretation of both objectives. In contrast to existing approaches, where the reward function is typically defined as a weighted sum of cost and comfort with heuristically selected weighting coefficients, the proposed approach enables a methodologically grounded balancing of these two objectives. To address the optimization problem, reinforcement learning (RL)-based methods are considered, including the Double Deep Q-Network (DDQN) and Proximal Policy Optimization (PPO), as well as approaches based on the use of pretrained Large Language Models (LLMs). At the time of writing this dissertation, no prior study was identified that investigates the application of LLMs in the context of household appliance scheduling in smart homes, nor their effectiveness in simulating decision-making processes analogous to those employed in RL. The proposed solutions are evaluated against a baseline knowledge-based system (KBS) defined by a human expert. The success criterion is established such that the considered solutions must achieve performance equal to or better than the baseline, while producing decisions that are justifiable from the perspective of a human expert. The results show that the PPO model satisfies the defined criteria and achieves the most stable and robust performance compared to the other considered approaches, whereas the DDQN model does not meet the established criteria. LLM-based approaches do not fully satisfy the defined criteria; however, they achieve performance comparable to the baseline, indicating their research potential.
--	--

	An important advantage of LLM-based approaches over RL methods lies in the elimination of the need for a time-consuming training process. It should also be noted that only base variants of these approaches are considered in this dissertation, leaving room for potential performance improvements. The results further demonstrate that the performance and strategies of the considered approaches strongly depend on the season of the year. In contrast to this dissertation, most existing studies do not examine the impact of seasonal variations. The obtained results additionally highlight an advantage of the proposed approaches over the baseline, reflected in their ability to automatically adapt to changing environmental conditions, whereas a human-defined KBS would require redesign under each significant change in the operating environment.
Date of endorsement by the scientific board:	20.11.2025.
Date of defence: (Filled in by the institution)	
Thesis defence board: (title, first name, last name, position, institution)	Chair: PhD, Milan Vidaković, Full Professor, Faculty of Technical Sciences, Novi Sad Member: PhD, Aleksandar Kovačević, Full Professor, Faculty of Technical Sciences, Novi Sad Member: PhD, Nikola Vojnović, Associate Professor, Faculty of Technical Sciences, Novi Sad Member: PhD, Miloš Radovanović, Full Professor, Faculty of Sciences, Novi Sad Mentor: PhD, Jelena Slivka, Full Professor, Faculty of Technical Sciences, Novi Sad PhD, Jovana Vidaković, Associate Professor, Faculty of Sciences, Novi Sad
Note:	

Zahvalnica

Želim da se zahvalim članovima komisije, čiji su komentari i smernice bili neophodni za uspešnu realizaciju ove disertacije.

Zahvaljujem se mentorki prof. dr. Jovani Vidaković na svim savetima i stručnoj podršci tokom izrade disertacije.

Zahvaljujem se mentorki prof. dr. Jeleni Slivki na razumevanju, prijateljskom pristupu i stručnim savetima tokom čitavog istraživačkog procesa i izrade disertacije.

Zahvaljujem se kolegama i prijateljima na razumevanju i podršci.

Želim da se zahvalim svojoj kumi Marini što je uvek bila tu da me sasluša i pruži neophodnu pozitivnu podršku.

Posebnu zahvalnost dugujem svojoj porodici – mami, tati i sestri Isidori – na podsticaju i ohrabrenju.

Konačno, zahvaljujem se svom suprugu Aleksandru na strpljenju i neiscrпноj podršci.

Ovu disertaciju posvećujem svojoj ćerki Anastasiji, koja me svakog dana inspiriše.

Indeks slika

Slika 1. Električna mreža evropskog domaćinstva (Vojnović et al. 2025). Svakom paru provodnika i tačke potrošnje (na primer, utičnica, svetiljka) dodeljen je jedinstveni identifikator u formatu: ID provodnika/ID tačke priključka.....	13
Slika 2. Pametno brojilo koje se sastoji od tri komponente: (1) komunikacionog modula, (2) pametnog brojila za merenje električne energije (izražene u kilovat-časovima) i (3) utičnice za povezivanje testiranog kućnog uređaja – elementarnog potrošača.	13
Slika 3. Višefazni Π_k segment (Vojnović et al. 2025).....	14
Slika 4. Dekuplovano kolo Γ_{ks} , $s \in a \vee b \vee c$ (Vojnović et al. 2025).....	15
Slika 5. Blok dijagram postupka sumiranja struja i korekcija napona.	16
Slika 6. Interakcija agenta i okruženja.....	18
Slika 7. Graf prelaza u Markovljevom procesu odlučivanja.....	19
Slika 8. Vizualizacija mehanizma odsecanja u PPO algoritmu za pozitivne i negativne vrednosti funkcije prednosti At . Leva strana prikazuje slučaj kada je $At > 0$: odsecanje ograničava rast ciljne funkcije za visoke vrednosti odnosa verovatnoća $pt\theta$, koje označavaju da trenutna politika favorizuje akciju at znatno više nego prethodna. Desna strana prikazuje slučaj kada je $At < 0$: ograničava se pad vrednosti funkcije za niske vrednosti $pt\theta$, koje ukazuju da trenutna politika daje znatno manju verovatnoću toj akciji u poređenju sa prethodnom politikom.	30
Slika 9. Na visokom nivou apstrakcije, veliki jezički model prima ulazni tekst u obliku prompta (eng. <i>prompt</i>) i, nakon obrade, generiše izlazni tekst (eng. <i>completion</i>).	34
Slika 10. Izlazni tekst se generiše iterativno, token po token. Prvi izlazni token se generiše na osnovu zadatog prompta. Svaki naredni token generiše se tako što model koristi prompt proširen prethodno generisanim tokenima kao novi ulaz.....	34
Slika 11. Tokenizator, na osnovu specifične metode i prethodno definisane procedure treniranja, segmentira tekst na tokene pre nego što se tekst preda LLM modelu. (OpenAI 2022).....	35
Slika 12. Sa leve strane prikazan je trenirani tokenizator, dok je sa desne strane prikazan LLM model koji sadrži embeddinge. Svaki token ima svoju odgovarajuću vektorsku reprezentaciju.	36
Slika 13. Transformer arhitektura (Vaswani et al. 2023).....	37
Slika 14. Pažnja zasnovana na skaliranom skalarnom proizvodu vektora.....	40
Slika 15. Višestruki mehanizam pažnje (Vaswani et al. 2023).....	42
Slika 16. Jednostavni prompt bez instrukcija gde LLM treba da dovrši započetu rečenicu.	44
Slika 17. Primer kompleksnog prompt koji sadrži više komponenti.....	45
Slika 18. Arhitektura sistema – komponente simulacionog okruženja za obuku i evaluaciju HEMS-a za upravljanje uređajima u pametnoj kući. U okviru sloja odluke moguće je odabrati jedan od sledećih pristupa: algoritam dvostruke duboke Q-mreže (DDQN), algoritam optimizacije proksimalne politike (PPO), jednu od dve varijante modela zasnovanog na velikom jezičkom modelu (LLM) ili model zasnovan na znanju (KBS).....	62
Slika 19. Primer dinamičkog određivanja značaja rada mašine za sudove.....	86
Slika 20. Prosečna nagrada po epizodi za zimski mesec (1. do 22. januar).....	93
Slika 21. Prosečna nagrada po epizodi za letnji mesec (1. do 22. jun).	93
Slika 22. Kontrolisanje unutrašnje temperature u toku zimskog dana.....	97
Slika 23. Upravljanje mašinom za sudove u zimskom danu.....	98
Slika 24. Upravljanje mašinom za veš u zimskom danu.....	98

Slika 25. Kontrolisanje unutrašnje temperature u toku letnjeg dana.	99
Slika 26. Upravlja mašinom za sudove u letnjem danu.	100
Slika 27. Upravljanje mašinom za veš u letnjem danu.	101
Slika 28. PPO model: agregiran uticaj karakteristika stanja kroz akcije u toku zimskog perioda.	104
Slika 29. DDQN model: agregiran uticaj karakteristika stanja kroz akcije tokom zimskog perioda.	105
Slika 30. PPO model: uticaj karakteristika stanja na akciju “uključi klima uređaj” tokom zimskog perioda.	107
Slika 31. PPO model: uticaj karakteristika na akciju “uključi klima uređaj” u <i>Typhoon HIL</i> simulacionom okruženju u zimskom danu za stanje u vremenskom koraku 07:15.	108
Slika 32. PPO model: upravljanje unutrašnjom temperaturom uz pregled proizvodnje energije putem solarnog panela i temperaturni komfor između 7:00 i 8:00 sati, u <i>Typhoon HIL</i> simulacionom okruženju.	108
Slika 33. PPO model: uticaj karakteristika na akciju “uključi klima uređaj” u <i>Typhoon HIL</i> simulacionom okruženju u zimskom danu za stanje u vremenskom koraku 00:00.	109
Slika 34. PPO model: upravljanje unutrašnjom temperaturom uz pregled proizvodnje energije putem solarnog panela i temperaturni komfor između 00:00 i 01:00 sati, u <i>Typhoon HIL</i> simulacionom okruženju.	109
Slika 35. DDQN model: uticaj karakteristika stanja na akciju “uključi klima uređaj” tokom zimskog perioda.	111
Slika 36. DDQN model: uticaj karakteristika na akciju “uključi klima uređaj” u <i>Load Flow</i> simulacionom okruženju u zimskom danu za stanje u vremenskom koraku 00:00.	111
Slika 37. DDQN model: upravljanje unutrašnjom temperaturom uz pregled proizvodnje energije putem solarnog panela i temperaturni komfor između 00:00 i 01:00 sati, u <i>Load Flow</i> simulacionom okruženju.	112
Slika 38. DDQN model: uticaj karakteristika na akciju “uključi klima uređaj” u <i>Load Flow</i> simulacionom okruženju u zimskom danu za stanje u vremenskom koraku 06:45.	112
Slika 39. DDQN model: upravljanje unutrašnjom temperaturom uz pregled proizvodnje energije putem solarnog panela i temperaturni komfor između 06:00 i 08:00 sati, u <i>Load Flow</i> simulacionom okruženju.	113
Slika 40. DDQN model: uticaj karakteristika na akciju “uključi klima uređaj” u <i>Load Flow</i> simulacionom okruženju u zimskom danu za stanje u vremenskom koraku 17:15.	113
Slika 41. DDQN model: upravljanje unutrašnjom temperaturom uz pregled proizvodnje energije putem solarnog panela i temperaturni komfor između 16:00 i 18:00 sati, u <i>Load Flow</i> simulacionom okruženju.	114
Slika 42. PPO model: uticaj karakteristika stanja na akciju “uključi mašinu za sudove” tokom zimskog perioda.	115
Slika 43. DDQN model: uticaj karakteristika stanja na akciju “uključi mašinu za sudove” tokom zimskog perioda.	116
Slika 44. DDQN model: Uticaj karakteristika na akciju “uključi mašinu za sudove” u <i>Typhoon HIL</i> simulacionom okruženju u zimskom danu za stanje u određenom vremenskom koraku.	117

Slika 45. DDQN model: Uticaj karakteristika na akciju “uključi mašinu za sudove” u <i>Load Flow</i> simulacionom okruženju u zimskom danu za stanje u određenom vremenskom koraku.	117
Slika 46. PPO model: uticaj karakteristika stanja na akciju “uključi veš mašinu” tokom zimskog perioda.....	118
Slika 47. DDQN model: uticaj karakteristika stanja na akciju “uključi veš mašinu” tokom zimskog perioda.....	119
Slika 48. DDQN model: Uticaj karakteristika na akciju “uključi mašinu za veš” u <i>Typhoon HIL</i> simulacionom okruženju u zimskom danu za stanje u određenom vremenskom koraku.	119
Slika 49. PPO model: uticaj karakteristika stanja agregirano kroz akcije u toku letnjeg perioda.	121
Slika 50. DDQN model: uticaj karakteristika stanja agregirano kroz akcije tokom letnjeg perioda.	121
Slika 51. PPO model: uticaj karakteristika stanja na akciju “uključi klima uređaj” tokom letnjeg perioda.	123
Slika 52. PPO model: Uticaj karakteristika na akciju “uključi klima uređaj” u <i>Typhoon HIL</i> simulacionom okruženju u letnjem danu za stanje u određenom vremenskom koraku.	123
Slika 53. PPO model: Uticaj karakteristika na akciju “uključi klima uređaj” u <i>Load Flow</i> simulacionom okruženju u letnjem danu za stanje u određenom vremenskom koraku.	123
Slika 54. DDQN model: uticaj karakteristika stanja na akciju “uključi klima uređaj” tokom letnjeg dana.....	124
Slika 55. PPO model: uticaj karakteristika stanja na akciju “uključi mašinu za sudove” tokom letnjeg perioda.	125
Slika 56. PPO model: Uticaj karakteristika na akciju “uključi mašinu za sudove” u <i>Typhoon HIL</i> simulacionom okruženju u letnjem danu za stanje u određenom vremenskom koraku.	126
Slika 57. DDQN model: uticaj karakteristika stanja na akciju “uključi mašinu za sudove” tokom letnjeg perioda.	127
Slika 58. DDQN model: Uticaj karakteristika na akciju “uključi mašinu za sudove” u <i>Typhoon HIL</i> simulacionom okruženju u letnjem danu za stanje u određenom vremenskom koraku.	128
Slika 59. DDQN model: Uticaj karakteristika na akciju “uključi mašinu za sudove” u <i>Load Flow</i> simulacionom okruženju u letnjem danu za stanje u određenom vremenskom koraku.	128
Slika 60. PPO model: uticaj karakteristika stanja na akciju “uključi veš mašinu” tokom letnjeg dana.....	129
Slika 61. PPO model: Uticaj karakteristika na akciju “uključi mašinu za veš” u <i>Typhoon HIL</i> simulacionom okruženju u letnjem danu za stanje u određenom vremenskom koraku.	129
Slika 62. PPO model: Uticaj karakteristika na akciju “uključi mašinu za veš” u <i>Load Flow</i> simulacionom okruženju u letnjem danu za stanje u određenom vremenskom koraku.	130
Slika 63. DDQN model: uticaj karakteristika stanja na akciju “uključi mašinu za veš” tokom letnjeg perioda.	131
Slika 64. DDQN model: Uticaj karakteristika na akciju “uključi mašinu za veš” u <i>Load Flow</i> simulacionom okruženju u letnjem danu za stanje u određenom vremenskom koraku.	131

Slika 65. Kontrolisanje unutrašnje temperature u toku zimskog dana.....	135
Slika 66. Udobnost u zimskom danu.	135
Slika 67. Upravljanje mašinom za sudove u zimskom danu.....	136
Slika 68. Upravljanje mašinom za veš u zimskom danu.....	137
Slika 69. Kontrolisanje unutrašnje temperature u toku letnjeg dana.	138
Slika 70. Upravljanje mašinom za sudove u analiziranom letnjem danu.....	138
Slika 71. Upravljanje mašinom za veš u letnjem danu.	139
Slika 72. Grafički prikaz performansi modela ostvarenih u <i>Load Flow</i> simulacionom okruženju u sedam uzastopnih testnih dana u januaru.	143
Slika 73. Grafički prikaz performansi modela ostvarenih u <i>Typhoon HIL</i> simulacionom okruženju u sedam uzastopnih testnih dana u januaru.	144
Slika 74. Grafički prikaz performansi modela ostvarenih u <i>Load Flow</i> simulacionom okruženju u sedam uzastopnih testnih dana u junu.....	145
Slika 75. Grafički prikaz performansi modela ostvarenih u <i>Typhoon HIL</i> simulacionom okruženju u sedam uzastopnih testnih dana u junu.....	145

Indeks tabela

Tabela 1. Primer MDP-a koji modeluje korišćenje električnog vozila: $ps's, a)$ i rs, a, s' predstavljaju verovatnoću prelaska iz stanja s u stanje s' preduzimanjem akcije a , tim redosledom. Neki prelazi su nemogući, te imaju verovatnoću 0 i nemaju dodeljenu nagrade.	19
Tabela 2. Uppedna analiza algoritama Q-učenje, DQN/DDQN i PPO	27
Tabela 3. Klasifikacija ponašanja ciljne funkcije $LCLIP\theta$ i ažuriranje parametara u zavisnosti od odnosa verovatnoća $pt\theta$ i funkcije prednosti At	30
Tabela 4. Primer izračunatih vrednosti pozicionog enkodinga (zaokruženo na dve decimalne).	38
Tabela 5. Vrednosti hiperparametara korišćene za obučavanje DDQN modela.	68
Tabela 6. Vrednosti hiperparametara korišćeni za obučavanje algoritma optimizacije proksimalne politike.	69
Tabela 7. Tekst prompta na osnovu koga je LLM generisao <i>Python</i> kod koji implementira sistem zasnovan na znanju za upravljanje uređajima pametne kuće.	70
Tabela 8. Sistemska poruka za usmeravanje LLM-a	76
Tabela 9. Uprošćen primer korisničkog prompta	78
Tabela 10. Dodatni tekst na kraju korisničkog prompta	79
Tabela 11. Tipovi uređaja koji su podržani u simulacionim okruženjima.	81
Tabela 12. Pregled tarifnih režima.	84
Tabela 13. Značaj rada uređaja u različitim periodima dana prema korisničkim zahtevima.	85
Tabela 14. Novčane vrednosti dodeljene radu uređaja na osnovu korisničkih prioriteta.	85
Tabela 15. Primer izračunavanja novčanog značaja rada klima uređaja tokom dana. Vrednost $BtAC$ predstavlja novčani značaj rada klima uređaja u vremenskom koraku t , u skladu sa opisom iz tabele 14. Vrednost $EtAC$ označava potrošnju električne energije klima uređaja u istom vremenskom koraku. Podaci o trenutnoj potrošnji preuzeti su iz simulacionog okruženja <i>Typhoon HIL</i> , opisanog u potpoglavlju 5.1.1.1. Komfort koji obezbeđuje klima uređaj modelovan je pomoću pojednostavljene jednačine (27), pri čemu se razmatra isključivo rad klima uređaja, bez penalizacije, uz pretpostavku da je potrošnja u svakom vremenskom koraku različita od nule ($EtAC \neq 0$).	87
Tabela 16. <i>Load Flow</i> okruženje: vrednosti nagrade ostvarene na testnom danu (23. dan januara i juna). Negativne vrednosti označavaju da je potrošnja energije bila veća od proizvodnje, dok pozitivne vrednosti označavaju da je proizvodnja električne energije veća od potrošnje.	95
Tabela 17. <i>Typhoon HIL</i> okruženje: vrednosti nagrade ostvarene na testnom danu (23. dan januara i juna). Negativne vrednosti označavaju da je potrošnja energije bila veća od proizvodnje, dok pozitivne vrednosti označavaju da je proizvodnja električne energije veća od potrošnje.	95
Tabela 18. <i>Load Flow</i> okruženje: vrednost funkcije nagrade ostvarene na testnim danima (23. januar i jun). Negativne vrednosti označavaju da je potrošnja energije bila veća od proizvodnje, dok pozitivne vrednosti označavaju da je proizvodnja električne energije veća od potrošnje.	133
Tabela 19. <i>Typhoon HIL</i> okruženje: vrednost funkcije nagrade ostvarene na testnim danima (23. januaru i junu). Negativne vrednosti označavaju da je potrošnja energije bila veća od proizvodnje, dok pozitivne vrednosti označavaju da je proizvodnja električne energije veća od potrošnje.	134

Tabela 20. Metrike performansi za zimski i letnji dan u <i>Load Flow</i> simulacionom okruženju, ukazujući na potrošnju električne energije, udobnost i ukupnu nagradu za sve uređaje zajedno, izražene u evrima. Negativne vrednosti označavaju da je potrošnja energije bila veća od proizvodnje, dok pozitivne vrednosti označavaju da je proizvodnja električne energije veća od potrošnje.....	140
Tabela 21. Metrike performansi za zimski i letnji dan u <i>Typhoon HIL</i> simulacionom okruženju, ukazujući na potrošnju električne energije, udobnost i ukupnu nagradu za sve uređaje zajedno, izražene u evrima. Negativne vrednosti označavaju da je potrošnja energije bila veća od proizvodnje, dok pozitivne vrednosti označavaju da je proizvodnja električne energije veća od potrošnje.....	141
Tabela 22. <i>Load Flow</i> okruženje: srednje vrednosti (<i>mean</i>) i standardne devijacije (<i>std</i>) metrika udobnosti, cene električne energije i ukupne nagrade, koje modeli ostvaruju tokom sedam testnih dana u januaru. Negativne vrednosti označavaju da je potrošnja energije bila veća od proizvodnje, dok pozitivne vrednosti označavaju da je proizvodnja električne energije veća od potrošnje.....	142
Tabela 23. <i>Typhoon HIL</i> okruženje: srednje vrednosti (<i>mean</i>) i standardne devijacije (<i>std</i>) metrika udobnosti, cene električne energije i ukupne nagrade, koje modeli ostvaruju tokom sedam testnih dana u januaru. Negativne vrednosti označavaju da je potrošnja energije bila veća od proizvodnje, dok pozitivne vrednosti označavaju da je proizvodnja električne energije veća od potrošnje.....	143
Tabela 24. <i>Load Flow</i> okruženje: srednje vrednosti (<i>mean</i>) i standardne devijacije (<i>std</i>) metrika udobnosti, cene električne energije i ukupne nagrade, koje modeli ostvaruju tokom sedam testnih dana u junu. Negativne vrednosti označavaju da je potrošnja energije bila veća od proizvodnje, dok pozitivne vrednosti označavaju da je proizvodnja električne energije veća od potrošnje.....	144
Tabela 25. <i>Typhoon HIL</i> okruženje: srednje vrednosti i standardne devijacije metrika udobnosti, cene električne energije i ukupne nagrade, koje modeli ostvaruju tokom sedam testnih dana u junu. Negativne vrednosti označavaju da je potrošnja energije bila veća od proizvodnje, dok pozitivne vrednosti označavaju da je proizvodnja električne energije veća od potrošnje.....	145

Sadržaj

Indeks slika	xiii
Indeks tabela	xvii
Rezime	xxiii
Abstract	xxv
1 Uvod	1
1.1 Značaj optimizacije potrošnje električne energije u pametnim kućama	1
1.2 Otvorena istraživačka pitanja u automatizaciji planiranja rada uređaja pametne kuće	2
1.3 Cilj, metodološki okvir i doprinosi istraživanja	4
1.4 Struktura disertacije	8
2 Teorijske osnove	11
2.1 Okruženja za simulaciju potrošnje električne energije u pametnim kućama ..	11
2.1.1 Proračun tokova snaga u kućnim instalacijama	12
2.1.1.1 Višefazni vodovi u domaćinstvu	13
2.1.1.2 Model i postupak tokova snaga kućnih mreža	14
2.2 Modeli dubokog učenja za upravljanje vremenom rada uređaja u simuliranom okruženju pametne kuće	16
2.2.1 Učenje potkrepljivanjem	16
2.2.1.1 Q-učenje	23
2.2.1.2 Duboka Q-mreža i dvostruka duboka Q-mreža	25
2.2.1.3 Proksimalna optimizacija politike	27
2.2.1.4 Određivanje uticaja karakteristika stanja na odluke modela učenja potkrepljivanjem	31
2.2.2 Veliki jezički modeli	32
2.2.2.1 Tokenizator	34
2.2.2.2 Transformer arhitektura	37
2.2.2.2.1 Pozicioni enkoding	37
2.2.2.2.2 Mehanizam pažnje	39
2.2.2.2.3 Enkoder i dekodeer	42
2.2.2.2.4 Strategije izbora sledećeg tokena	43
2.2.2.3 Tehnika prompt inženjeringa	43
2.3 Analiza odluka modela zasnovanih na učenju potkrepljivanjem	46
3 Pregled aktuelnog stanja u oblasti	49
3.1 Istorijski pregled razvoja sistema za upravljanje energijom u pametnim kućama	49
3.2 Savremene strategije učenja potkrepljivanjem za upravljanje energijom u pametnim kućama	50

3.3	Primena velikih jezičkih modela u sistemima za upravljanje energijom u pametnim kućama.....	57
3.4	Ograničenja postojećih rešenja i obrazloženje potrebe za daljim istraživanjem.....	58
4	Metodologija.....	61
4.1	Arhitektura sistema.....	61
4.2	Dizajn adaptivnog simulacionog okruženja.....	62
4.3	Dizajn funkcije nagrade.....	63
4.4	Dizajn modela učenja potkrepljivanjem za određivanje optimalnog vremena rada uređaja pametne kuće.....	66
4.4.1	Postavka hiperparametara za model dvostruke duboke Q-mreže.....	67
4.4.2	Postavka hiperparametara algoritma optimizacije proksimalne politike.....	68
4.5	Dizajn prompta koji omogućava velikom jezičkom modelu da odredi optimalno vreme rada uređaja pametne kuće.....	70
4.5.1	Zero-shot strategija prompta.....	70
4.5.2	Few-shot strategija prompta.....	75
5	Evaluacija metodologije.....	81
5.1	Studija slučaja: eksperimentalna postavka za simulaciju pametne kuće.....	81
5.1.1	Modelovanje uređaja u pametnoj kući.....	81
5.1.1.1	Klima uređaj.....	82
5.1.1.2	Mašina za sudove.....	83
5.1.1.3	Mašina za veš.....	83
5.1.1.4	Solarni panel.....	84
5.1.2	Vremenski uslovi i tarifni režimi.....	84
5.1.3	Definisanje zahteva stanara pametne kuće.....	84
5.1.4	Reprezentacija stanja i skup mogućih akcija.....	87
5.1.4.1	Reprezentacije stanja.....	87
5.1.4.2	Skup akcija.....	88
5.1.5	Eksperimentalna postavka.....	88
6	Rezultati i diskusija.....	91
6.1	Analiza rezultata modela učenja potkrepljivanjem.....	91
6.1.1	Analiza procesa konvergencije nagrade.....	91
6.1.2	Analiza vrednosti funkcije nagrade na testnom danu koji sledi neposredno nakon perioda treniranja.....	94
6.1.3	Analiza odluka modela na nivou dana.....	96
6.1.3.1	Analiza rada uređaja u zimskom danu.....	96
6.1.3.2	Analiza rada uređaja u letnjem danu.....	99
6.1.3.3	Zaključak analize odluka model na nivou zimskog i letnjeg dana..	101

6.1.4	Analiza uticaja karakteristika stanja na odluke modela	102
6.1.4.1	Interpretacija odluka	102
6.1.4.1.1	Odluke u zimskom periodu.....	103
6.1.4.1.1.1	Zaključak na nivou odluka u zimskom periodu	119
6.1.4.1.2	Odluke u letnjem periodu	120
6.1.4.1.2.1	Zaključak na nivou odluka u letnjem periodu	131
6.1.4.1.3	Zaključak interpretacije odluka na nivou različitih godišnjih doba.....	132
6.2	Analiza rezultata velikog jezičkog modela	133
6.2.1	Analiza vrednosti funkcije nagrade na nivou simulacionog okruženja.....	133
6.2.2	Analiza odluka modela na nivou dana	134
6.2.2.1	Analiza rada uređaja u zimskom danu	134
6.2.2.2	Analiza rada uređaja u letnjem danu.....	137
6.2.2.3	Zaključak analize odluka model na nivou zimskog i letnjeg dana ..	139
6.3	Poređenje modela na nivou jednog dana sa KBS definisanim od strane ljudskog eksperta	139
6.4	Ispitivanje robusnosti modela	141
6.4.1	Zaključak ispitivanja robusnosti modela.....	146
6.5	Zaključci i potencijalna poboljšanja	148
7	Zaključak.....	151
	Literatura	155
	Biografija.....	167

Pametne kuće postaju sve prisutnije u savremenom životu. Primena obnovljivih izvora energije i mogućnosti skladištenja energije pružaju potencijal za ekonomičniju i održiviju potrošnju električne energije u domaćinstvima. Međutim, manuelno upravljanje uređajima u ovakvim sistemima zahteva vreme, kao i određeni nivo tehničkog znanja, zbog čega bi automatizacija sistema za upravljanje uređajima mogla značajno da rastereti stanare. U ovoj disertaciji razmatra se problem automatskog upravljanja uređajima u pametnoj kući, sa ciljem istovremene optimizacije potrošnje električne energije i udobnosti stanara.

Razvijeno je modularno simulaciono okruženje koje omogućava realističnu simulaciju rada kućnih uređaja i jednostavno definisanje različitih konfiguracija pametnih kuća i eksternih uslova u kojima se one nalaze. Implementirana su dva pristupa simulaciji rada uređaja: simulacija sa konstantnom potrošnjom energije tokom rada uređaja i simulacija sa varijabilnom potrošnjom energije u okviru radnog ciklusa, čime se omogućava verodostojnije predstavljanje realnih sistema. Prethodna istraživanja su se pretežno oslanjala na simulacije u kojima je potrošnja uređaja modelovana kao konstantna tokom njihovog rada. Međutim, rezultati ove disertacije ukazuju da izbor načina modelovanja potrošnje uređaja može imati značajan uticaj na performanse razmatranih pristupa za optimizaciju rada kućnih uređaja.

Definisana je višeciljna funkcija nagrade koja objedinjuje troškove električne energije i komfor stanara. Komfor stanara je modelovan u monetarnim jedinicama i predstavlja meru njihove spremnosti da plate korišćenje kućnih uređaja u određenim vremenskim intervalima. Na ovaj način omogućeno je direktno objedinjavanje komfora i troškova električne energije u okviru jedinstvene funkcije nagrade, uz jasno ekonomsko tumačenje oba kriterijuma. Za razliku od postojećih pristupa, u kojima se funkcija nagrade najčešće definiše kao ponderisana suma troškova i komfora sa heuristički odabranim težinskim koeficijentima, predloženi pristup omogućava metodološki utemeljeno balansiranje ova dva cilja.

Kao rešenja za optimizaciju postavljenog problema razmatrani su modeli zasnovani na učenju potkrepljivanjem, uključujući dvostruku duboku Q-mrežu i optimizaciju proksimalne politike, kao i pristupi zasnovani na upotrebi pretreniranih velikih jezičkih modela (VJM). U trenutku pisanja ove disertacije nije pronađena nijedna studija koja je razmatrala primenu VJM u kontekstu optimizacije vremena rada kućnih uređaja, niti njihovu efikasnost u simulaciji procesa donošenja odluka analognih onima koji se koriste u učenju potkrepljivanjem.

Predložena rešenja upoređena su sa referentnim modelom zasnovanim na znanju definisanom od strane ljudskog eksperta. Kao kriterijum uspeha postavljeno je da razmatrana rešenja ostvare iste ili bolje performanse u odnosu na referentni model, uz donošenje odluka koje su opravdane iz perspektive ljudskog eksperta. Rezultati pokazuju da model zasnovan na optimizaciji proksimalne politike ispunjava postavljene kriterijume i ostvaruje najstabilnije i najrobusnije performanse u poređenju sa ostalim razmatranim pristupima, dok dvostruka duboka Q-mreža ne ispunjava postavljene kriterijume. Pristupi zasnovani na velikim jezičkim modelima (VJM) nisu u potpunosti zadovoljili postavljene kriterijume, ali su ostvarili performanse uporedive sa referentnim

modelom, što ukazuje na njihov istraživački potencijal. Prednost pristupa zasnovanih na VJM u odnosu na metode učenja potkrepljivanjem ogleda se u eliminaciji potrebe za dugotrajnim procesom treniranja. Takođe, treba istaći da su u okviru ove disertacije razmatrane osnovne varijante ovih pristupa, te da postoji prostor za unapređenje njihovih performansi.

Rezultati su pokazali da performanse i strategije razmatranih pristupa u velikoj meri zavise od godišnjeg doba. Za razliku od ove disertacije, većina postojećih radova ne razmatra uticaj sezonskih razlika. Dobijeni rezultati dodatno ukazuju na prednost predloženih pristupa u odnosu na referentni model, koja se ogleda u njihovoj sposobnosti automatske adaptacije na promenljive uslove okruženja, dok bi sistem zasnovan na znanju, definisan od strane ljudskog eksperta, zahtevao redizajn pri svakoj značajnijoj promeni okruženja.

Abstract

Smart homes are becoming increasingly prevalent in everyday life. The integration of renewable energy sources and energy storage systems provides significant potential for more economical and sustainable electricity consumption in smart homes. However, manual control of appliances in such systems requires time as well as a certain level of technical expertise, which suggests that automation of appliance management systems could substantially reduce the burden on residents. This dissertation addresses the problem of automatic appliance control in a smart home, with the objective of simultaneously optimizing electricity consumption and occupant comfort.

A modular simulation environment has been developed, enabling realistic simulation of household appliances as well as straightforward definition of various smart home configurations and external environmental conditions. Two approaches to appliance consumption modeling are implemented: simulation with constant energy consumption during appliance operation and simulation with variable energy consumption within the operating cycle, allowing for a more faithful representation of real-world systems. Previous studies have predominantly relied on simulations in which appliance energy consumption is modeled as constant throughout operation. However, the results of this dissertation indicate that the choice of appliance consumption modeling approach can have a significant impact on the performance of the considered methods for appliance operation optimization.

A multi-objective reward function is defined that jointly accounts for electricity costs and occupant comfort. Occupant comfort is modeled in monetary terms and represents a measure of residents' willingness to pay for the operation of household appliances during specific time intervals. This formulation enables the direct integration of comfort and electricity costs within a unified reward function, providing a clear economic interpretation of both objectives. In contrast to existing approaches, where the reward function is typically defined as a weighted sum of cost and comfort with heuristically selected weighting coefficients, the proposed approach enables a methodologically grounded balancing of these two objectives.

To address the optimization problem, reinforcement learning (RL)-based methods are considered, including the Double Deep Q-Network (DDQN) and Proximal Policy Optimization (PPO), as well as approaches based on the use of pretrained Large Language Models (LLMs). At the time of writing this dissertation, no prior study was identified that investigates the application of LLMs in the context of household appliance scheduling in smart homes, nor their effectiveness in simulating decision-making processes analogous to those employed in RL.

The proposed solutions are evaluated against a baseline knowledge-based system (KBS) defined by a human expert. The success criterion is established such that the considered solutions must achieve performance equal to or better than the baseline, while producing decisions that are justifiable from the perspective of a human expert. The results show that the PPO model satisfies the defined criteria and achieves the most stable and robust performance compared to the other considered approaches, whereas the DDQN model does not meet the established criteria. LLM-based approaches do not fully satisfy the defined criteria; however, they achieve performance comparable to the baseline,

indicating their research potential. An important advantage of LLM-based approaches over RL methods lies in the elimination of the need for a time-consuming training process. It should also be noted that only base variants of these approaches are considered in this dissertation, leaving room for potential performance improvements.

The results further demonstrate that the performance and strategies of the considered approaches strongly depend on the season of the year. In contrast to this dissertation, most existing studies do not examine the impact of seasonal variations. The obtained results additionally highlight an advantage of the proposed approaches over the baseline, reflected in their ability to automatically adapt to changing environmental conditions, whereas a human-defined KBS would require redesign under each significant change in the operating environment.

1 Uvod

Uvodno poglavlje podeljeno je na pet celina. Potpoglavlje 1.1 definiše problem optimizacije potrošnje električne energije uz očuvanje komfora stanara u pametnim kućama i prikazuje motive za njegovo rešavanje. Potpoglavlje 1.2 ističe otvorena istraživačka pitanja i razmatra različite pristupe rešavanju ovog problema. Potpoglavlje 1.3 precizira predmet istraživanja, njegove ciljeve i naglašava doprinos ove disertacije. Konačno, potpoglavlje 1.4 pruža sažet pregled strukture disertacije.

1.1 Značaj optimizacije potrošnje električne energije u pametnim kućama

Razvoj obnovljivih i održivih izvora energije postao je globalni prioritet. Solarna energija, kao najbrže rastući obnovljivi izvor, široko se primenjuje u pametnim kućama za smanjenje zavisnosti od tradicionalne električne mreže (Chehri and Mouftah 2013). Međutim, kapacitet ovog izvora energije je promenljiv zbog njegove velike zavisnosti od meteoroloških uslova (Maimó-Far et al. 2020). Usled toga su domaćinstva ipak ovisna o električnoj mreži, što je pogotovo izraženo u momentu vrhunca potrošnje, kada stanari dolaze kući sa posla, a solarna energija je u padu (Kostková et al. 2013) (Zhang et al. 2013). Električne mreže pokušavaju da balansiraju fluktuacije potrošnje električne energije primenom tarifnih inicijativa koje podstiču domaćinstva da optimizuju svoju potrošnju (Kirschen et al. 2000). Osim toga, domaćinstva značajno doprinose godišnjim emisijama ugljen-dioksida (Fenner et al. 2018), te bi optimizacija njihove potrošnje kroz optimalno iskorišćenje obnovljivih izvora energije imala značajan pozitivan efekat na životnu sredinu (Ahmad et al. 2017).

Optimizacija potrošnje električne energije u pametnim kućama može se postići strateškim planiranjem vremena rada uređaja čiji rad može biti odložen bez značajnog uticaja na komfor stanara (Agnetis et al. 2013). Na primer, stanari mogu uključiti takve uređaje tokom perioda niže tarife električne energije ili kada je proizvodnja energije iz solarnog panela visoka. Međutim, manuelno kreiranje rasporeda rada uređaja predstavlja vremenski zahtevan i složen zadatak koji nije intuitivan (Barbato et al. 2009). Zbog složenosti i varijabilnosti u obrascima potrošnje i proizvodnje energije, automatizacija planiranja vremena rada uređaja postaje neophodna (Raza et al. 2024). Sistemi za upravljanje energijom u domaćinstvu (eng. *House Energy Management Systems*, HEMS) dizajnirani su da rešavaju brojne izazove, uključujući automatizaciju planiranja vremena rada uređaja, kako bi uskladili suprotstavljene ciljeve: minimizaciju troškova električne energije i očuvanje komfora stanara (Han et al. 2023).

Uzimajući u obzir značaj automatizacije rada kućnih aparata za optimizaciju potrošnje energije uz očuvanje komfora stanara, u ovoj disertaciji predložena je primena tehnika dubokog učenja za rešavanje ovog problema. Razlozi za odabir ovog predmeta istraživanja su višestruki.

Motiv za rešavanje navedenog problema proizlazi iz kontinuiranog rasta potražnje za energijom kao ključnog faktora ekonomskog razvoja država, pri čemu sektor energetike ostaje dominantan izvor emisije gasova sa efektom staklene bašte (eng. *greenhouse*

gases). Najveći doprinos emisijama dolazi iz sektora proizvodnje električne energije i grejanja, koji su 2021. godine generisali 16 milijardi tona ekvivalenta ugljen-dioksida (Ritchie et al. 2020). U okviru ciljeva održivog razvoja (eng. *Sustainable Development Goals*, SDG), SDG 7 naglašava potrebu za obezbeđivanjem pristupačne, pouzdane, održive i moderne energije za sve, dok SDG 13 ističe hitnost preduzimanja mera za suzbijanje klimatskih promena i njihovih posledica (Barbier and Burgess 2017). Stoga je neophodno intenzivno promovisati obnovljive izvore energije, poput solarne energije, kako bi se smanjile emisije ugljen-dioksida i ublažile posledice globalnog zagrevanja (Allegretti et al. 2022). Integracija obnovljivih izvora energije u pametne kuće može značajno doprineti smanjenju emisija ugljen-dioksida na globalnom nivou. Na primer, energetski efikasna porodična kuća u Poljskoj, opremljena solarnim panelima, generisala je 87,5% manje emisija ugljen-dioksida u poređenju sa tradicionalnom kućom zagrevanom ugljem (Siudek et al. 2020).

Motiv za primenu tehnika dubokog učenja u rešavanju ovog problema proizilazi iz učestale upotrebe arhitekture neuronskih mreža sa propagacijom unapred (eng. *feedforward neural network*) u optimizaciji potrošnje električne energije u pametnim kućama (Leitao et al. 2020). Dalji tehnološki napredak, uključujući razvoj metoda dubokog učenja potkrepljivanjem (eng. *Deep Reinforcement Learning*, DRL) omogućio je efikasnije rešavanje složenih problema u ovoj oblasti. Pregled literature ukazuje da je duboko učenje potkrepljivanjem trenutno najsavremeniji (eng. *state-of-the-art*) pristup za automatizaciju rada kućnih uređaja, s ciljem optimizacije potrošnje električne energije uz očuvanje komfora stanovnika. Otvorena istraživačka pitanja u ovoj oblasti detaljnije su analizirana u potpoglavlju 1.2.

S obzirom na složenost problema optimizacije potrošnje električne energije, ovo istraživanje fokusira se na orkestraciju rada kućnih uređaja na nivou pojedinačnog domaćinstva. Međutim, predloženi pristup može se prilagoditi i za šire energetske sisteme, uključujući stambene zgrade i celokupne urbane četvrti.

1.2 Otvorena istraživačka pitanja u automatizaciji planiranja rada uređaja pametne kuće

Jedan od ključnih izazova u implementaciji HEMS za automatizaciju planiranja rada uređaja jeste razvoj ciljne funkcije koja balansira dva suprotstavljena cilja: smanjenje troškova električne energije i povećanje komfora stanara. Za ovakve više-ciljne optimizacione probleme, posebno su se pokazale efikasnim tehnike ponderisanog zbira i *Pareto Front* (Gomes et al. 2022). Kombinovanje učenja potkrepljivanjem sa tehnikom ponderisanog zbira predstavlja najsavremeniji pristup u implementaciji HEMS sistema (Aldahmashi and Ma 2024) (Pan et al. 2024), zbog čega je tehnika ponderisanog zbira usvojena pri definisanju ciljne funkcije u ovom istraživanju. Ipak, kao što je prikazano u potpoglavlju 3.2 formulacija ciljne funkcije nije jednoznačno određena ni u okviru ovog pristupa, te i dalje ostaje otvoreno istraživačko pitanje.

Još jedan značajan izazov u razvoju HEMS-a jeste kreiranje okruženja za simulaciju pametne kuće. Simulacije su ključne, budući da dizajniranje i evaluacija planova rada uređaja u stvarnim domovima nisu praktični. Razvoj rešenja zahteva testiranje različitih planova rada uređaja, pri čemu neki mogu biti neefikasni, dovesti do povećanih troškova električne energije i narušiti navike i komfor stanara. Pregled literature koji su sproveli *Gomes* i saradnici (Gomes et al. 2022) pokazao je da istraživači koriste različite

simulacione postavke za pametne kuće, pri čemu je *Matlab* najčešće korišćen kao simulaciono okruženje. Međutim, nedostatak standardizovanog simulacionog okruženja otežava pouzdano poređenje razvijenih rešenja. Jedna od ključnih razlika među postojećim simulacijama odnosi se na pristup modelovanju potrošnje uređaja u različitim vremenskim koracima tokom njihovog rada – bilo kao konstantnu ili promenljivu vrednost. Autori radova (Gholami and Sanjari 2021), kao i (Esmael Nezhad et al. 2021), definišu potrošnju uređaja u svakom vremenskom koraku kao konstantu vrednost određenu na osnovu srednje ili maksimalne potrošnje uređaja tokom celokupnog radnog ciklusa. Ovaj pristup pojednostavljuje realnu situaciju, budući da određeni uređaji, poput veš mašine, imaju značajne varijacije u potrošnji energije tokom različitih radnih faza. Kako bi povećali realističnost simulacije, autori rada (Lee and Choi 2019) uvode model sa različitim nivoima potrošnje za klima uređaj, pored standardnog pristupa modelovanja sa konstantnom maksimalnom potrošnjom.

Kada su ciljna funkcija i okruženje za simulaciju pametne kuće dizajnirani, implementacija HEMS može se ostvariti primenom različitih tehnika. Tradicionalni pristupi uključuju dinamičko programiranje (Zhao and Keerthisinghe 2020) i stohastičko programiranje (Zheng et al. 2021), kao i heurističke i meta-heurističke tehnike poput genetskih algoritma (Gutierrez-Martinez et al. 2019) i optimizacije rojem čestica (Rezaee Jordehi 2020). Takođe se koriste pristupi kao što su fazi logika (Kontogiannis et al. 2021), konvolucione neuronske mreže (Chellaswamy et al. 2021) i sistemi zasnovani na znanju (eng. *Knowledge-Based Systems*, KBS) (X. Zhou et al. 2023). U kontekstu HEMS, primena učenja potkrepljivanjem predstavlja najmoderniju tehniku (Gomes et al. 2022). Ova tehnika je posebno pogodna za HEMS zbog svoje sposobnosti da se prilagođava dinamičkim okruženjima kroz kontinuirane interakcije, što eliminiše potrebu za eksplisnim programiranjem svakog mogućeg scenarija (Shakya et al. 2023).

Učenje potkrepljivanjem i sistemi zasnovani na znanju su trenutno najpopularniji pristupi za automatizaciju planiranja vremena rada uređaja u pametnim kućama (Al-Ani and Das 2022). Ipak, u okviru HEMS-a, oba pristupa imaju određena ograničenja. Okruženje pametne kuće se često menja usled sezonskih temperaturnih varijacija, promena u skupu i karakteristikama uređaja pametne kuće ili izmena u tarifnoj politici električne mreže.

Trenutno najuspešnija tehnika iz oblasti učenja potkrepljivanjem za implementaciju HEMS-a je proksimalna optimizacija politike (eng. *Proximal Policy Optimization*, PPO). PPO algoritam je dizajniran za okruženja sa nepoznatom dinamikom, što ga čini posebno pogodnim za rešavanje HEMS problema. Međutim, PPO tipično zahteva veliki broj uzoraka za efikasno obučavanje (Arulkumaran et al. 2017) i postaje suboptimalan kada dođe do promene u okruženju, što zahteva ponovno obučavanje (Zhou et al. 2016).

Slično tome, implementacija sistema zasnovanih na znanju zahteva angažovanje ljudskih stručnjaka koji bi morali redizajnirati ponašanje sistema za svaki mogući scenario (Raza et al. 2024). Ovaj pristup nije skalabilan jer bi svaka izmena u okruženju zahtevala ponovno prilagođavanje sistema zasnovanog na znanju.

Po znanju autora, mali broj studija koja predlaže HEMS rešenje zasnovano na učenju potkrepljivanjem razmatrao je interpretabilnost rezultujućih modela. Ovakva analiza mogla bi doprineti boljem razumevanju odluka modela, povećati poverenje korisnika u sistem i olakšati njegovo dalje usavršavanje (Beechey et al. 2023). Stoga se u ovom istraživanju predlaže analiza uticaja karakteristika stanja na odluke modela učenja potkrepljivanjem korišćenjem SHAP (eng. *Shapley Additive Explanations*) biblioteke

(Lundberg and Lee 2017), koja pruža matematičke metode za kvantifikaciju doprinosa pojedinačnih karakteristika stanja u konačnim predikcijama obučanih modela (Merrick and Taly 2020).

Primena unapred obučanih velikih jezičkih modela (eng. *Large Language Models*, LLM) mogla bi doprineti ublažavanju ograničenja rešenja zasnovanih na učenju potkrepljivanjem i sistemima zasnovanih na znanju. Za razliku od ovih pristupa, LLM-ovi imaju potencijal da se prilagode promenama u podacima bez potrebe za opsežnim ponovnim treniranjem ili finim podešavanjem specifičnim za domen (Cao et al. 2024). Oslanjajući se na velike količine postojećeg znanja, LLM-ovi mogu smanjiti zavisnost od stručnjaka iz određenog domena i ubrzati proces razvoja (Kaplan et al. 2020). Njihova sposobnost da se nose sa raznovrsnim i složenim scenarijima čini ih potencijalno pogodnim za primenu u dinamičnim i nepredvidivim okruženjima (Gao et al. 2023).

Postojeća istraživanja o upotrebi LLM-ova u kontekstu HEMS-a su pretežno usmerena na interaktivno savetovanje stanovnika o metodama za poboljšanje komfora i smanjenje potrošnje električne energije (potpoglavlje 3.3). U ovim pristupima, kontrola uređaja ostaje manuelna, pri čemu korisnici donose konačne odluke. Po znanju autora, ne postoji istraživanje koje se bavi korišćenjem LLM-ova za automatizaciju upravljanja uređajima u cilju rešavanja definisanog više-ciljnog problema.

U ovom istraživanju uvodi se i evaluira hipoteza da se LLM-ovi mogu efikasno primeniti u implementaciji HEMS-a koji balansira minimizaciju troškova električne energije i održavanje komfora stanara pametne kuće kroz automatizaciju planiranja vremena rada uređaja. Ova hipoteza zasniva se na sposobnost LLM-ova da prepoznaju obrasce i logički rezonuju u složenim sekvencama podataka (Brown et al. 2020). Takve karakteristike doprinele su ohrabrujućim rezultatima koje su LLM-ovi ostvarili u predviđanju vremenskih serija (eng. *time series*) (Gruver et al. 2023) (Jin et al. 2024.).

1.3 Cilj, metodološki okvir i doprinosi istraživanja

Cilj istraživanja je razvoj metodologije za automatizaciju planiranja vremena rada uređaja u pametnoj kući radi smanjenja troškova električne energije uz održanje komfora stanara. Ovaj cilj se razlaže na sledeće podciljeve istraživanja:

- C₁ Dizajn simulacionog okruženja koje verno oponaša potrošnju električne energije u pametnoj kući. Simulaciono okruženje treba da bude adaptivno, što podrazumeva mogućnost definisanja razolikih realističnih scenarija po pitanju pametnih kuća sa različitim tipovima uređaja³ i uslovima u kojima se one nalaze – vremenskim uslovima i tarifnim režimima, koji zavise od geografske lokacije.
- C₂ Definisanje metrike za evaluaciju komfora stanara pametne kuće i načina na koji se ta metrika kombinuje sa cenom električne energije. Metrika komfora bi trebala da omogućiti stanovnicima da na intuitivan način izraze svoje želje vezane za rad uređaja i kompromis tih želja sa uštedom električne energije. Definisana metrika bi imala ulogu funkcije nagrade koja bi se koristila za evaluaciju rešenja.

³ Primeri tipova uređaja su proizvođači energije, skladišta energije i standardni potrošači. U slučaju standardnih potrošača, biće razmotrena mogućnost simulacije uređaja sa varijabilnom potrošnjom energije.

- C₃ Dizajn rešenja za automatizaciju planiranja vremena rada uređaja u pametnoj kući s ciljem postizanja maksimalne definisane nagrade. Analiza literature (poglavlje 3) ukazuje na potrebu razmatranja primene metoda učenja potkrepljivanjem i velikih jezičkih modela pri razvoju rešenja. Kao osnovni referentni pristup (eng. *baseline*) neophodno je usvojiti manuelno definisan sistem zasnovan na znanju.
- C₄ Evaluacija razvijenog rešenja za automatizaciju planiranja vremena rada uređaja korišćenjem definisane funkcije nagrade i dizajniranog simulacionog okruženja. Ključno je da evaluacija obuhvati različita godišnja doba, s obzirom na njihov uticaj na dinamiku potrošnje električne energije u pametnoj kući (potpoglavlje 3.2).
- C₅ Analiza prednosti i nedostataka razvijenog rešenja sprovođenjem analize objašnjivosti razvijenog rešenja.

Pretpostavke na kojima se zasniva predloženo istraživanje su sledeće:

- P₁ Adaptivno okruženje za simulaciju pametne kuće dovoljno verno simulira realističnu pametnu kuću da može biti iskorišćeno za pouzdanu evaluaciju rešenja.
- P₂ Moguće je definisati metriku komfora stanara i njen kompromis sa potrošnjom električne energije tako da ona bude intuitivna stanovnicima.
- P₄¹ Definisana metrika za evaluaciju rešenja zaista meri zadovoljstvo stanovnika razvijenim rešenjem.
- P₄² Definisana metrika za evaluaciju omogućava rangiranje rešenja po njihovom kvalitetu.
- P₅ Moguće je izvršiti analizu objašnjivosti modela koja bi omogućila analizu njegovih prednosti i nedostataka.

Hipoteza koja će biti evaluirana u istraživanju je da su metode učenja potkrepljivanjem i veliki jezički modeli efikasni pristupi za rešavanje postavljenog optimizacionog problema. Efikasnost se definiše kao postizanje iste ili bolje nagrade u odnosu na manuelno definisan sistem zasnovan na znanju, uz to da su odluke rešenja adekvatne iz perspektive ljudskih eksperata. Adekvatnost podrazumeva da ljudski ekspert, kroz analizu objašnjivosti modela, može da potvrdi da se odluke koje donosi model u datoj situaciji razumne.

U nastavku je dat sažetak metodologije koja je sprovedena, dodatno opisana i u (Kaplar et al. 2024), radi postizanja prethodno definisanih ciljeva.

Prvi korak istraživanja obuhvatao je **implementaciju simulacionog okruženja**. U tu svrhu razvijen je sistem koji se sastoji iz četiri sloja:

1. Fizički sloj – Simulira potrošnju energije uređaja u pametnoj kući. U istraživanju razmatrana su dva pristupa modelovanju potrošnje energije:
 - Potrošnja uređaja modelovana je kao konstantna vrednost tokom celokupnog trajanja rada uređaja (Typhoon HIL 2008).
 - Potrošnja uređaja modelovana je kao varijabilna vrednost, zavisna od faze rada uređaja (Vojnović et al. 2022) (Vojnović et al. 2023b) (Vojnović et al. 2023a).

2. Sloj za donošenje odluka (HEMS) – Orkestrira rad uređaja sa ciljem optimizacije potrošnje električne energije uz očuvanje komfora stanara.
3. Informacioni sloj – Obuhvata istorijske podatke o vremenskim uslovima i tarifnim režimima dostupnim na određenoj geografskoj lokaciji, koji se koriste za simulaciju.
4. Komunikacioni sloj – Omogućava razmenu informacija između fizičkog sloja i sloja za donošenje odluka.

Sledeći korak istraživanja obuhvatao je **definisanje funkcije nagrade**. U tu svrhu analizirana je relevantna literatura koja predlaže matematičke modele za optimizaciju energetske efikasnosti i korisničkog komfora u HEMS sistemima zasnovanim na učenju potkrepljivanjem i velikim jezičkim modelima. Na osnovu ove analize, funkcija nagrade definisana je kao ponderisana suma troškova električne energije i kvantifikacije komfora stanovnika. Troškovi električne energije određeni su u zavisnosti od tarifnih režima u različitim vremenskim intervalima, kako je prikazano u potpoglavlju 5.1.2. Stanari izražavaju svoje zahteve u vezi sa komforom kroz odabir jednog od tri nivoa značaja rada uređaja u određenom vremenskim intervalima. Kako bi se kvantifikacija komfora mogla sabrati sa troškovima električne energije, nivoi značaja izraženi su u monetarnim jedinicama, pri čemu je značaj rada uređaja povezan sa tarifnim režimima, kako je opisano u potpoglavlju 5.1.3.

Treći korak istraživanja bio je **razvoj HEMS pristupa** koji će biti korišćeni u sloju za donošenje odluka. Analizom literature, utvrđeno je da treba razmotriti tri pristupa razvoju HEMS-a:

1. Učenje potkrepljivanjem. Razmatrana su dva algoritma: dvostruka duboka Q-mreža (eng. *Double Deep Q-Network*, DDQN) i proksimalna optimizacija politike (eng. *Proximal Policy Optimization*, PPO).
2. Primena pretreniranih velikih jezičkih modela. U okviru ovog pristupa, evaluirane su dve strategije:
 - *Zero-shot* strategija – LLM-u je prosleđena instrukcija da generiše HEMS sistem zasnovan na pravilima. Rezultujući HEMS je sličan sistemu zasnovanom na ručno definisanim pravilima, s tom razlikom što pravila ne formuliše čovek, već ih LLM generiše na osnovu znanja stečenog tokom obuke modela.
 - *Few-shot* strategija – LLM je tretiran kao model učenja potkrepljivanjem. U svakom promptu, model je primao opis trenutnog stanja i birao optimalnu akciju. Pri tome, na raspolaganju su mu bili primeri prethodnih interakcija, gde su za prethodna stanja i sprovedene akcije bile dostupne vrednosti neposredne nagrade.
3. Sistem zasnovan na ručno definisanim pravilima, definisan od strane autora rada.

Četvrti korak, **evaluacija metodologije**, sprovedena je u dve faze – zimskoj i letnjoj sezoni, uzimajući u obzir njihov uticaj na dinamiku potrošnje električne energije u pametnoj kući. Ovaj pristup omogućio je realističniju procenu performansi razvijenih modela u različitim uslovima rada. Detalji eksperimentalne postavke su pojašnjeni u potpoglavlju 5.1.5.

Konačno, sprovedena je **analiza rezultata**, koja je obuhvatila:

1. Evaluaciju performansi modela na osnovu definisane funkcije nagrade, pri čemu su analizirane pojedinačne metrike komfora, potrošnje energije, kao i njihova kombinacija.
2. Ručnu analizu odluka modela putem vizuelizacije njegovih odluka tokom jednog nasumično odabranog dana iz letnje i jednog iz zimske sezone. Ova analiza omogućila je uvid u ponašanje modela u specifičnim situacijama.
3. Analizu modela učenja potkrepljivanjem, gde su vizualizovane dnevne prosečne nagrade postignute tokom treniranja, radi procene konvergencije modela. Dodatno, za ove modele sprovedena je SHAP analiza, kojom je procenjen doprinos pojedinačnih karakteristika stanja u konačnim predikcijama obučениh modela.

U ovoj disertaciji predstavljeni su novi pristupi i alati za optimizaciju potrošnje električne energije uz očuvanje komfora stanara. Doprinosi teze su sledeći:

- **Adaptivno okruženje za simulaciju pametne kuće:** Razvijeno je adaptivno okruženje koje omogućava simulaciju raznolikih realističnih scenarija potrošnje energije u pametnim kućama. Simulaciono okruženje omogućava stanaru da kreira virtuelni model svog doma, usklađen sa stvarnim uslovima. Na osnovu tako konfigurisane simulacije, modeli se mogu obučavati u bezbednom i kontrolisanom okruženju, što omogućava procenu njihovih performansi pre primene u praksi. Nakon procesa obučavanja, modeli, koji poseduju visok stepen prilagođenosti specifičnim uslovima stanara, mogu biti implementirani u realnom okruženju uz manju adaptaciju na razlike između simulacije i realnog okruženja. Time se umanjuju gubici koji bi nastali u početnim fazama treniranja modela i skraćuje ukupno vreme neophodno za treniranje modela, zahvaljujući bržem izvođenju simulacija u odnosu na realno vreme.
- **Predlog funkcije nagrade za minimizaciju troškova električne energije uz očuvanje komfora stanara:** Metrika komfora stanara izražena je u monetarnim jedinicama, čime je omogućeno njeno direktno sabiranje sa troškovima električne energije. Na ovaj način kvantifikuje se spremnost stanara da plate korišćenje određenih uređaja u definisanom vremenskom periodu. Stanari definišu nivo prioriteta rada uređaja (visok, srednji i nizak), pri čemu se svakom nivou dodeljuje određena monetarna vrednost, usklađena sa tarifnim režimima (visoka, srednja i niska tarifa). Za razliku od predloženog pristupa, u postojećoj literaturi funkcija nagrade se najčešće formulisala kao ponderisana suma troškova električne energije i komfora, pri čemu su težinski koeficijenti često bili birani heuristički, bez jasnog metodološkog objašnjenja.
- **Način modelovanja potrošnje uređaja:** U disertaciji su razmatrana dva pristupa modelovanju potrošnje električnih uređaja: model sa konstantnom potrošnjom energije tokom rada i model sa varijabilnom potrošnjom energije. Varijabilni režim potrošnje verodostojnije oslikava ponašanje uređaja u realnim sistemima, čime se trenirani model čini pogodnijim za primenu u stvarnim uslovima eksploatacije. Nasuprot tome, u značajnom delu postojeće literature uređaji se modeluju pod pretpostavkom konstantne potrošnje tokom celokupnog radnog ciklusa, što predstavlja pojednostavljenje realnog procesa. U praksi, obrasci potrošnje električnih uređaja su složeni i obuhvataju više faza rada sa različitim nivoima potrošnje energije. Rezultati ove disertacije ukazuju da izbor načina modelovanja potrošnje uređaja ima značajan uticaj na performanse predloženih rešenja za upravljanje radom električnih uređaja.

- **Utvrđena je potreba za periodičnim ažuriranjem modela u skladu sa sezonskim promenama spoljnih uslova:** Analiza je pokazala da modeli donose značajno različite odluke u zavisnosti od sezone, što ukazuje na ograničenja modela treniranih i evaluiranih na vremenski usko ograničenim podacima. Većina postojećih radova ne razmatra sezonske razlike tokom procesa treniranja i evaluacije. U praktičnoj primeni spoljašnji faktori se menjaju tokom vremena, zbog čega je neophodno da se modeli periodično ponovo obučavaju korišćenjem proširenog skupa podataka koji uključuje najnovije dostupne informacije. Na ovaj način omogućava se kontinuirana adaptacija modela na sezonske promene i dugoročne obrasce u potrošnji i proizvodnji električne energije. U produkcionim sistemima, pad vrednosti funkcije nagrade može se koristiti kao indikator potrebe za ažuriranjem modela novom, adaptiranom verzijom.
- **Procena mogućnosti primene LLM na zadati problem:** Sve veći broj istraživanja razmatra potencijal primene LLM-ova kao sistema za prepoznavanje obrazaca u različitim domenima, uključujući i modelovanje nejezičkih obrazaca (Mirchandani et al. 2023). Međutim, u trenutku pisanja disertacije, nije postojala nijedna studija koja je istražila kapacitet LLM-ova da modeluju šablone i donose odluke u kontekstu rada uređaja u pametnim kućama. Takođe, nisu pronađene studije koje su istraživale efikasnost pretreniranih LLM-ova u simulaciji procesa donošenja odluka sličnog onom koji se koristi u učenju potkrepljivanjem. Zahvaljujući ogromnom ugrađenom znanju, LLM-ovi se mogu brzo prilagoditi promenama u podacima bez skupog ponovnog treniranja, čime uklanja potrebu RL-a za čestim ponovnim obučavanjem.

1.4 Struktura disertacije

Disertacija je struktuirana u sedam poglavlja, koja detaljno pokrivaju sve aspekte predloženog istraživanja.

U **drugom poglavlju** predstavljena je teorijska osnova neophodna za razumevanje sprovedenog istraživanja. Opisani su simulacioni sistemi pametnih kuća koji su korišćeni, pri čemu je poseban naglasak stavljen na razliku između *Typhoon HIL* i *Load Flow* simulacionih okruženja. Takođe, istaknuta je prednost *Load Flow* okruženja u kontekstu optimizacije i simulacije energetske efikasnosti. Nakon toga, definisani su osnovni koncepti modela učenja potkrepljivanjem i velikih jezičkih modela, čime se postavlja teorijski okvir za njihovu primenu u upravljanju energijom u pametnim kućama.

Treće poglavlje pruža pregled aktuelnog stanja u oblasti upravljanja energijom u pametnim kućama. U njemu je dat istorijski pregled razvoja ove oblasti, opisane su savremene strategije zasnovane na učenju potkrepljivanjem, kao i strategije primene velikih jezičkih modela u optimizaciji potrošnje energije. Na kraju poglavlja opisani su ograničenja i nedostaci postojećih rešenja kao i potreba za detaljnijom analizom.

Četvrto poglavlje detaljno opisuje metodologiju razvoja predloženog sistema. Prikazana je arhitektura HEMS-a, koja obuhvata simulaciono okruženje, modele za donošenje odluka i bazu eksternih podataka. Posebna pažnja posvećena je definisanju funkcije nagrade korišćene tokom obuke i evaluacije modela. Konačno, predstavljen je dizajn modela zasnovanih na učenju potkrepljivanjem i promptova za velike jezičke modele korišćenih u rešavanju problema automatskog određivanja vremena rada uređaja u pametnoj kući.

Peto poglavlje bavi se eksperimentalnim postupkom sprovedenim radi evaluacije razvijenog rešenja. Definisana je eksperimentalna postavka za simulaciju pametne kuće, koja uključuje modelovanje uređaja, vremenskih uslova i tarifa, definisanje zahteva stanara, reprezentaciju stanja i definisanje skupa mogućih akcija. Detaljno su prikazani rezultati rada uređaja čiji je rad automatizovan od strane modela dubokog učenja i velikog jezičkog modela. Rezultati su grupisani na osnovu godišnjeg doba i tipa simulacionog okruženja (*Typhoon HIL* i *Load Flow*). Takođe je sprovedeno ispitivanje robusnosti modela kroz procenu njihove sposobnosti da održe stabilne performanse u danima nakon treniranja, uprkos promenljivim spoljašnjim faktorima.

Šesto poglavlje prikazuje rezultate i diskusiju modela učenja potkrepljivanjem i velikih jezičkih modela primenjenih u upravljanju pametnom kućom. Analizirani su proces konvergencije, vrednosti funkcije nagrade i odluke modela na nivou dana, uz interpretaciju doprinosa pojedinačnih karakteristika pomoću SHAP metoda. Rezultati LLM pristupa upoređeni su sa sistemom zasnovanim na pravilima definisanim od strane eksperta. Na kraju je ispitana robusnost modela u različitim simulacionim okruženjima i predložena potencijalna poboljšanja razvijenih rešenja.

Sedmo poglavlje sadrži zaključak i sažetak najvažnijih rezultata, ističući ograničenja istraživanja i moguću primenu realizovanog sistema i modela u praksi. Takođe su dati mogući pravci budućih istraživanja.

2 Teorijske osnove

Ovo poglavlje izlaže teorijske osnove istraživanja, neophodne za razumevanje metodologije korišćene u ovoj disertaciji. U potpoglavlju 2.1 predstavljeni su osnovni principi rada okruženja za simulaciju potrošnje električne energije u pametnim kućama. Potpoglavlje 2.2 predstavlja ključne koncepte modela dubokog učenja primenjenih u automatizaciji planiranja vremena rada uređaja u pametnim kućama. Konkretno, potpoglavlje 2.2.1 prezentuje osnovne pojmove učenja potkrepljivanjem, dok potpoglavlje 2.2.2 sažima ključne koncepte velikih jezičkih modela.

2.1 Okruženja za simulaciju potrošnje električne energije u pametnim kućama

Dizajn sistema za automatizaciju plana rada uređaja u pametnoj kući zahteva poređenje više alternativnih rešenja. Evaluacija u stvarnim domovima nije praktična, jer neoptimalna rešenja mogu povećati troškove električne energije i narušiti komfor stanara. Pored toga, testiranje u realnim uslovima trajalo bi više meseci, što bi onemogućilo brzu iteraciju i unapređenje rešenja. Stoga je ključno razviti simulaciono okruženje koje omogućava brzu i preciznu evaluaciju različitih pristupa. U ovom istraživanju analizirana je primena dva simulaciona okruženja potrošnje električne energije u pametnim kućama: *Typhoon HIL* (Typhoon HIL 2008) i *Load Flow* (Vojnović et al. 2022) (Vojnović et al. 2023b) (Vojnović et al. 2023a).

Radi pojednostavljenja simulacije, mnoga istraživanja (Forootani et al. 2022), (Huang et al. 2022), (Li et al. 2020), (Aldahmashi and Ma 2024) i (Pan et al. 2024) modeluju potrošnju uređaja kao konstantnu tokom celog perioda njihovog rada. Međutim, u realnim uslovima, potrošnja električne energije uređaja poput mašine za pranje veša, mašine za pranje sudova i šporeta varira u zavisnosti od faze rada (Pipattanasomporn et al. 2014). Iako je modelovanje potrošnje kao konstantne metodološki jednostavnije, ono pruža manje realističnu simulaciju dinamike rada uređaja u poređenju sa pristupima koji uzimaju u obzir varijabilnost potrošnje. Sa druge strane, u radovima (Xu et al. 2021), (Xu et al. 2020), i (Liu et al. 2020) korišćena su simulaciona okruženja koja omogućavaju preciznije modelovanje potrošnje uređaja u zavisnosti od operativne faze. U ovom istraživanju su isprobana oba pristupa – modelovanje potrošnje kao konstantne realizovano je korišćenjem platforme *Typhoon HIL*, dok je modelovanje varijabilne potrošnje uređaja realizovano upotrebom *Load Flow* simulacionog okruženja.

Typhoon HIL je komercijalno simulaciono okruženje koje je u okviru ove disertacije korišćeno za modelovanje uređaja u domaćinstvu. Ovo okruženje omogućava precizno simuliranje električnih sistema i ponašanja uređaja u različitim uslovima. Detalji implementacije nisu prikazani, budući da komercijalna priroda sistema ne omogućava uvida u sve tehničke aspekte njegove realizacije. U okviru ovog istraživanja, *Typhoon HIL* je korišćen za simulaciju uređaja sa konstantnom potrošnjom energije po vremenskom koraku. Simulaciono okruženje *Load Flow* omogućava modelovanje promenljive potrošnje i detaljnije je prikazano u nastavku poglavlja.

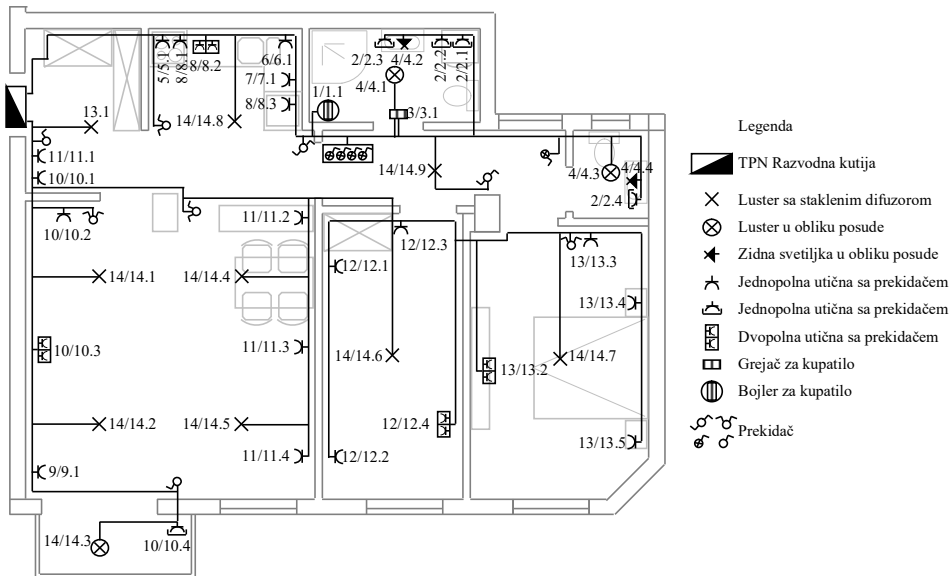
Temperatura unutar kuće, koja je ključan faktor komfora stanara, kao i rad određenih uređaja, poput klima uređaja i solarnih panela, zavise od spoljnih vremenskih uslova. Kako simulacija bila što realističnija, prikupljeni su istorijski podaci o vremenskim uslovima za određenu geografsku lokaciju tokom različitih perioda godine. Po ugledu na radove (Huang et al. 2022), (Li et al. 2020), (Liu et al. 2020), (Lissa et al. 2021), ovi podaci su integrisani u simulaciju, omogućavajući verodostojnije modelovanje rada uređaja poput solarnih panela i klima uređaja.

2.1.1 Proračun tokova snaga u kućnim instalacijama

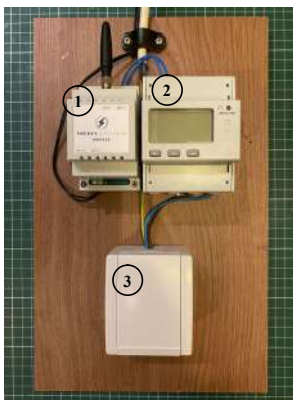
U ovoj disertaciji modelovana je električna mreža evropskog domaćinstva, prikazana na slici 1 (Vojnović et al. 2025). Električna mreža se sastoji od četrnaest glavnih strujnih kola, koja se napajaju preko trofazne razvodne table (eng. *Three-Phase and Neutral distribution board*, TPN).

Za potrebe modelovanja potrošača, merena je potrošnja električne energije u realnim uslovima. Merenje je izvršeno pomoću uređaja koji će, u okviru ove disertacije, biti nazvan pametno brojilo (slika 2). Ovaj uređaj sastoji se iz sledećih komponenti: (1) komunikacionog modula, (2) pametnog brojila za merenje električne energije (izražene u kilovat-časovima), i (3) utičnice za povezivanje testiranog kućnog uređaja. Komunikacioni modul koristi MODBUS protokol (Thomas 2008) za prikupljanje podataka sa pametnog brojila i njihovo slanje u oblak putem Wi-Fi veze. Izmereni podaci koriste se za modelovanje potrošača, dok se izmerene električne karakteristike i dužine grana provodnika koriste za modelovanje segmenata električnih instalacija.

Radi kreiranja detaljnog modela potrošača, sprovedeno je merenje u trajanju od 24 sata, što je ključno za uređaje sa vremenski promenljivim profilima potrošnje električne energije (na primer, mašina za pranje sudova). Tokom celog radnog ciklusa uređaja, kao što je mašina za veš, zabeležena je precizna promena potrošnje, uključujući sve faze rada (pranje, ispiranje, centrifugiranje, sušenje). Snimljeni vremenski zavisni profili potrošnje korišćeni su u simulacionom okruženju tako što bi se, po aktivaciji uređaja, reprodukovao čitav prethodno zabeleženi profil potrošnje. Na taj način, simulacija verno odražava realne obrasce potrošnje u situacijama kada je uređaj aktivan.



Slika 1. Električna mreža evropskog domaćinstva (Vojnović et al. 2025). Svakom paru provodnika i tačke potrošnje (na primer, utičnica, svetiljka) dodeljen je jedinstveni identifikator u formatu: ID provodnika/ID tačke priključka.



Slika 2. Pametno brojilo koje se sastoji od tri komponente: (1) komunikacionog modula, (2) pametnog brojlara za merenje električne energije (izražene u kilovat-časovima) i (3) utičnice za povezivanje testiranog kućnog uređaja – elementarnog potrošača.

2.1.1.1 Višefazni vodovi u domaćinstvu

Električne instalacije u evropskim domaćinstvima najčešće obuhvataju četvorožilne i petožilne trofazne konfiguracije ($3\phi 4w$, $3\phi 5w$), kao i jednofazne trožične konfiguracije ($1\phi 3w$). Uzemljeni provodnik dodatno je inkorporiran u vod radi bezbednosti i održavanja nultog potencijala, ali je neaktivan tokom normalnog režima rada.

Karsonove jednačine (eng. *Carson's equation*), uz Kronovu redukciju (eng. *Kron reduction*), integrišu impedansu zemlje u model sistema. Na slici 3 prikazan je generalni Π_k segment koji predstavlja višefaznu konfiguraciju električnih instalacija u kući (Vojnović et al. 2025). Redna matrica $\hat{\mathbf{Z}}_k^{abc}$, prikazana na slici 3, obuhvata impedanse koje predstavljaju i sopstvenu i međusobnu induktivnost između provodnika. Matrice

otočnih admitansi $\hat{\mathbf{Y}}_k^{abc}$ i $\hat{\mathbf{Y}}_k^{\prime abc}$ (eng. *shunt*) modeluju sopstvene i međusobne kapacitivne sprege, koje se u praksi obično tretiraju kao identične. Serijske i otočne matrice su dimenzija $s \times s$, gde s označava broj faza. Model vodova u višefaznim domaćinstvima detaljno je opisano sledećim jednačinama (Vojnović et al. 2025):

$$\hat{\mathbf{U}}_k^{abc} = \hat{\mathbf{U}}_i^{abc} - \hat{\mathbf{Z}}_k^{abc} \hat{\mathbf{I}}_k^{abc}, \quad (1a)$$

$$\hat{\mathbf{I}}_{k0}^{abc} = \hat{\mathbf{Y}}_k^{abc} \hat{\mathbf{U}}_i^{abc}, \hat{\mathbf{I}}_{k0}^{\prime abc} = \hat{\mathbf{Y}}_k^{\prime abc} \hat{\mathbf{U}}_k^{abc}, \quad (1b)$$

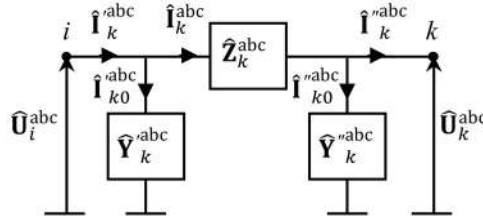
$$\hat{\mathbf{I}}_k^{abc} = \hat{\mathbf{I}}_k^{abc} + \hat{\mathbf{I}}_{k0}^{abc}, \hat{\mathbf{I}}_k^{\prime abc} = \hat{\mathbf{I}}_{k0}^{\prime abc} + \hat{\mathbf{I}}_k^{abc} + \hat{\mathbf{I}}_k^{\prime abc}. \quad (1c)$$

gde su (slika 3): (i) vektori napona čvorova $\hat{\mathbf{U}}_i^{abc}$, $\hat{\mathbf{U}}_k^{abc}$ vektori napona čvorova (ii) vektori struje rednih grana $\hat{\mathbf{I}}_k^{abc}$, $\hat{\mathbf{I}}_k^{\prime abc}$, $\hat{\mathbf{I}}_{k0}^{abc}$ vektori struje rednih grana, i (iii) vektori struja otočnih grana $\hat{\mathbf{I}}_{k0}^{abc}$, $\hat{\mathbf{I}}_{k0}^{\prime abc}$ vektori struja otočnih grana. Navedeni vektori su istih dimenzija $s \times 1$. Za trofazni segment kućne instalacije, redne i otočne matrice date su sledećim izrazima:

$$\hat{\mathbf{Z}}_k^{abc} = \begin{bmatrix} \hat{Z}_{kaa} & \hat{Z}_{kab} & \hat{Z}_{kac} \\ \hat{Z}_{kab} & \hat{Z}_{kbb} & \hat{Z}_{kbc} \\ \hat{Z}_{kca} & \hat{Z}_{kcb} & \hat{Z}_{kcc} \end{bmatrix}, \quad (2)$$

$$\hat{\mathbf{Y}}_k^{abc} = \hat{\mathbf{Y}}_k^{\prime abc} = \begin{bmatrix} \hat{Y}_{kaa} & \hat{Y}_{kab} & \hat{Y}_{kac} \\ \hat{Y}_{kab} & \hat{Y}_{kbb} & \hat{Y}_{kbc} \\ \hat{Y}_{kca} & \hat{Y}_{kcb} & \hat{Y}_{kcc} \end{bmatrix}.$$

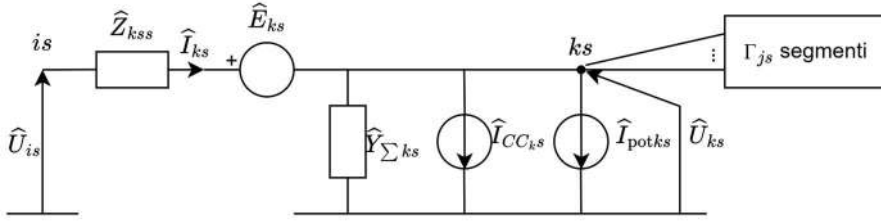
Iz ovih matrica su isključeni redovi i kolone koje se odnose na nedostajuće faze u višefaznom segmentu kućne mreže.



Slika 3. Višefazni Π_k segment (Vojnović et al. 2025).

2.1.1.2 Model i postupak tokova snaga kućnih mreža

Uzimajući u obzir različite konfiguracije vodova kućnih instalacija za potrebe proračuna tokova snaga ($1\phi 3w$) i ($3\phi 4w$), izvedeno je deкупlovano kolo Γ_{ks} , gde $s \in \phi_k$ označava fazu segmenta (slika 4), a ϕ_k označava skup faza iz segmenata Π_k (slika 3). Kolo Γ_{ks} može biti povezano sa više segmenata naslednika (Γ_{js} sa slike 4), pri čemu se indeksi ovih naslednika čuvaju u skupu naslednika α_{ks} . Segment Γ_{ks} sadrži čvorove is i ks , $s \in \phi_k$.



Slika 4. Dekuplovano kolo Γ_{ks} , $s \in a \vee b \vee c$ (Vojnović et al. 2025).

Impedansa $\hat{Z}_{kss}$ ($s \in a \vee b \vee c$) iz relacije (2) predstavlja sopstvenu induktivnost segmenta k i odgovara dijagonalnom elementu matrice $\hat{\mathbf{Z}}_k^{abc}$ sa slike 3. Admitansa $\hat{Y}_{\Sigma ks}$ predstavlja ukupnu otočnu admitansu segmenta Π_k i njihovih segmenata Π_j naslednika. Ove admitanse predstavljaju sopstvene kapacitivnosi segmenta k .

Ukupna admitansa se izračunava sabiranjem admitansi prethodnog segmenta Π_k i sledećih segmenata Π_j . Kapacitivne sprege sa drugim fazama modeluju se strujnim izvorom $\hat{I}_{CCks}$, dok se induktivne sprege sa drugim fazama označava sa $\hat{E}_{ks}$. Naponi čvorova is i ks , $s \in \phi_k$ označeni su redom $\hat{U}_{is}$ i $\hat{U}_{ks}$. Struja redne grane označena je sa $\hat{I}_{ks}$, dok $\hat{I}_{potks}$ predstavlja struju potrošača. Potrošač je modelovan preko kompleksne snage $\hat{S}_{ks}$.

Postupak proračuna tokova snaga sastoji se iz dva koraka (Vojnović et al. 2025):

1. Korak na gore – sumiranje struja grana (eng. *Backward sweep*):

$$(\hat{I}_{potks})^{h+1} = \frac{[\hat{S}_{ks}(\hat{U}_{ks}^h)]}{(\hat{U}_{ks}^h)^*}, \quad (2a)$$

$$(\hat{I}_{CCks})^{h+1} = \sum_{t \in \phi_k \wedge t \neq s} \hat{Y}_{kst}(\hat{U}_{kt})^{h+1}, \quad (2b)$$

$$(\hat{I}_{ks})^{h+1} = \hat{Y}_{\Sigma ks}(\hat{U}_{ks})^{h+1} + (\hat{I}_{CCks})^{h+1} + (\hat{I}_{conks})^{h+1} + \sum_{t \in \alpha_{ks}} (\hat{I}_{ts})^{h+1}, \quad (2c)$$

$s = a, b, c;$

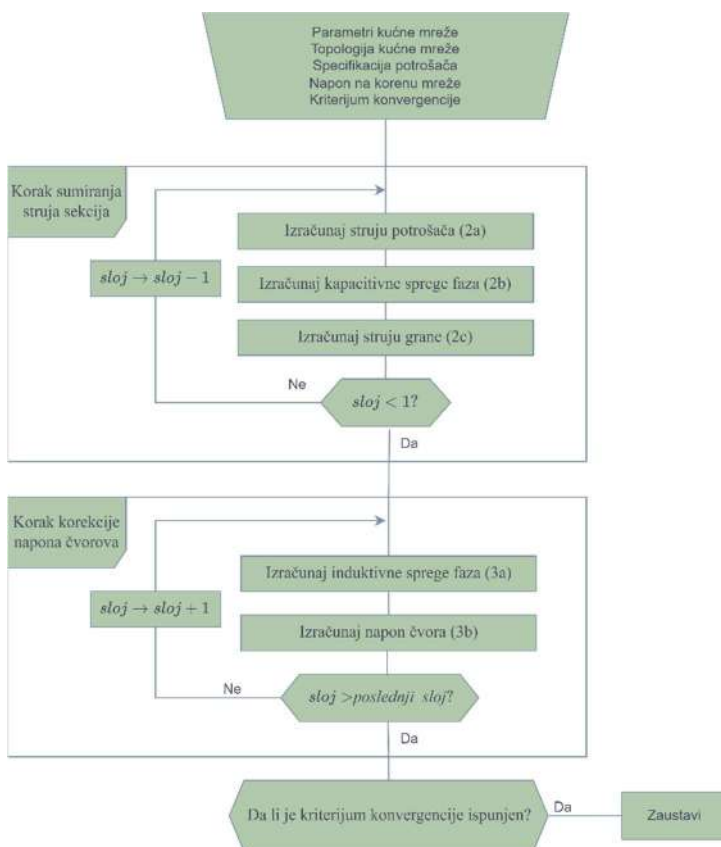
2. Korak korekcije napona čvorova (eng. *Forward sweep*):

$$(\hat{E}_{ks})^{h+1} = \sum_{t \in \phi_k \wedge t \neq s} \hat{Z}_{kst}(\hat{I}_{kt})^{h+1}, \quad (3a)$$

$$(\hat{U}_{is})^{h+1} = (\hat{U}_{ks})^{h+1} - \hat{Z}_{kss}(\hat{I}_{ks})^{h+1} - (\hat{E}_{ks})^{h+1}, \quad s = a, b, c. \quad (3b)$$

Slika 5 prikazuje blok-dijagram postupka sumiranja struja grana i korekcije napona čvorova (eng. *Backward-Forward Sweep*, BFS). Postupak se završava kada su ispunjeni kriterijumi konvergencije.

U predloženom postupku petlje se ne obrađuju, jer su električne instalacije u domaćinstvima strogo radialne strukture. U distributivnim slaboupetljanim mrežama, petlje se rešavaju korišćenjem *Tevnenove teoreme* (Vojnović et al. 2023b).



Slika 5. Blok dijagram postupka sumiranja struja i korekcija napona.

2.2 Modeli dubokog učenja za upravljanje vremenom rada uređaja u simuliranom okruženju pametne kuće

U ovoj disertaciji razmatra se primena modela dubokog učenja (eng. *deep learning*) za upravljanje vremenom rada uređaja u pametnoj kući. Konkretno, razmatrani su modeli učenja potkrepljivanjem (eng. *deep reinforcement learning*) i veliki jezički modeli (eng. *large language models*). Potpoglavlje 2.2.1 prikazuje osnove učenja potkrepljivanjem, sa fokusom na modele dvostrukih dubokih Q-mreža (eng. *double deep Q-network*) (Hasselt et al. 2015) i modele zasnovane na proksimalnoj optimizaciji politike (eng. *proximal policy optimization*) (Schulman et al. 2017), koji predstavljaju najsavremenija rešenja za primenu učenja potkrepljivanjem u problemu upravljanja vremenom rada uređaja u pametnoj kući. Potpoglavlje 2.2.2 predstavlja osnovne koncepte velikih jezičkih modela, koji su u ovoj disertaciji razmatrani kao alternativa naprednim tehnikama učenja potkrepljivanjem.

2.2.1 Učenje potkrepljivanjem

Za razliku od nadgledanog mašinskog učenja (eng. *supervised learning*), gde se model obučava na označenim skupovima podataka, učenje potkrepljivanjem (eng.

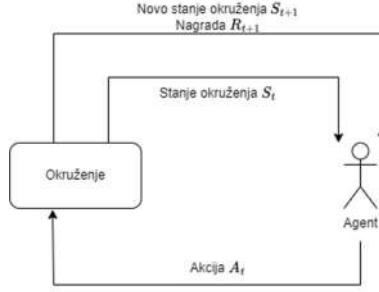
reinforcement learning) zahteva okruženje u kojem model može posmatrati trenutna stanja okruženja (eng. *state*), isprobavati moguće akcije u datom stanju i analizirati njihove ishode, čime postepeno unapređuje svoje performanse kroz interakciju sa okruženjem (Gupta et al. 2021). Učenje potkrepljivanjem se primenjuje kada je neophodno rešiti problem kroz niz uzastopnih akcija, pri čemu njihov zajednički efekat dovodi do konačnog rešenja (Nikolic and Zecevic 2019). Ključna pretpostavka ovog pristupa (i osnovna razlika u odnosu na nadgledano učenje) jeste da model nema unapred dostupne informacije o tome koje su akcije bile ispravne u datom kontekstu, a koje ne.

Markovljevi procesi odlučivanja (eng. *Markov Decision Process*, MDP) su matematička osnova za reprezentaciju algoritama učenja potkrepljivanjem. MDP opisuje sistem koji se sastoji od tri glavna elementa:

- *Agent* (eng. *agent*), čiji je cilj da nauči koju akciju treba preduzeti u kontekstu datog stanja okruženja.
- *Okruženje* (eng. *environment*), koje predstavlja sistem koji agent posmatra i u kom deluje. Okruženje definiše konačan skup stanja S u kojima se može naći agent i dopustivih akcija $A(s)$ koje agent može da preduzme u konkretnim stanjima. Stanje okruženja predstavlja specifične informacije koje agent opaža i koje koristi prilikom donošenja odluka o preduzimanju akcija.
- *Nagrada* (eng. *reward*) je numerička vrednost koja ocenjuje akciju agenta u datom stanju okruženja. Cilj agenta je da postigne najveću dugoročnu nagradu (eng. *cumulative reward*) – ukupnu očekivanu sumu nagrada koje agent može akumulirati kroz vreme kao rezultat pojedinačnih nagrada.

Slika 6 ilustruje interakciju između okruženja i agenta koji interaguju u diskretnim vremenskim trenucima. U svakom vremenskom trenutku t , agent opaža reprezentaciju stanja okruženja, $S_t \in S$, na osnovu koje preduzima akciju, $A_t \in A$. Kao posledicu ove akcije, dobija neposrednu numeričku nagradu (eng. *immediate reward*) $R_{t+1} \in R \subset \mathbb{R}$ i prelazi u novo stanje okruženja S_{t+1} . Ishod učenja predstavlja *optimalnu politiku*, odnosno, pravilo preslikavanja stanja okruženja u odgovarajuće akcije koje dovodi do maksimalne dugoročne nagrade (Nikolic and Zecevic 2019).

Interakcija između agenta i okruženja može se posmatrati kao sekvenca akcija. Ako se sekvenca akcija uvek završava konačnim stanjem (eng. *terminal state*) u konačnom broju akcija, tada ovu sekvencu zovemo epizodom, a zadatak epizodnim zadatkom (eng. *episodic task*). Epizode su međusobno nezavisne, odnosno početak epizode ne zavisi od toga kako se prethodna epizoda završila. Međutim, u mnogim realnim scenarijima ne postoji jasno definisano konačno stanje, pa se takvi zadaci nazivaju neograničeni (eng. *continuing task*).



Slika 6. Interakcija agenta i okruženja

U konačnom (eng. *finite*) MDP-u, skupovi stanja, akcija i nagrada (S, A i R) imaju konačan broj elemenata. U ovom slučaju, slučajne promenljive R_t i S_t imaju diskretne raspodele verovatnoće koje zavise isključivo od prethodnog stanja i akcije. Funkcija prelaska stanja p , koja određuje verovatnoće (Pr) prelaska iz jednog stanja s u novo stanje s' preduzimanjem akcije a , koja rezultuje nagradom r je definisana na sledeći način (Nikolic and Zecevic 2019):

$$p(s', r | s, a) = \Pr\{S_{t+1} = s', R_{t+1} = r \mid S_t = s, A_t = a\}, \quad (4)$$

za sve $s', s \in S, r \in R$ i $a \in A(s)$. U MDP, funkcija prelaska stanja p u potpunosti karakteriše dinamiku okruženja.

Za MDP je ključno Markovljevo svojstvo (eng. *Markov property*), koje definiše da novo stanje S_{t+1} i nagrada R_{t+1} u vremenskom trenutku $t + 1$ zavise isključivo od prethodnog stanja okruženja S_t i preduzete akcije A_t u vremenskom trenutku t , a ne od celokupne istorije procesa donošenja odluka. Modelovanje MDP-a je prikazano u primeru 2.1.

Primer 2.1 Proces korišćenja električnog vozila (eng. *Electric Vehicle*, EV) može se modelovati kao MDP. EV poseduje bateriju čije se stanje napunjenosti izražava u procentima, pri čemu vrednost varira u opsegu od 0% do 100%. Umesto precizne numeričke reprezentacije napunjenosti baterije, moguće je koristiti pojednostavljen skup stanja $S = \{low, high\}$, gde stanje *low* označavaju napunjenost baterije ispod 50%, dok stanje *high* označava napunjenost baterije iznad 50%. U svakom stanju, agent može odabrati jednu od sledećih akcija: da ne preduzme ništa (eng. *idle*), da puni bateriju (eng. *recharge*) ili da je prazni (eng. *discharge*). Skup dopustivih akcija zavisi od trenutnog stanja baterije:

- (1) ako je $S = low$, dostupne akcije su $A(low) = \{idle, discharge, recharge\}$
- (2) ako je $S = high$, dostupne akcije su $A(high) = \{idle, discharge\}$, pri čemu punjenje baterije nije dozvoljeno.

Tabela 1 definiše ovaj MDP – verovatnoće prelaza između stanja i odgovarajuće nagrade za svaku akciju, dok slika 7 ilustruje graf prelaza između stanja. Prilikom primene akcije *discharge* iz stanja *high*, postoje dva moguća ishoda:

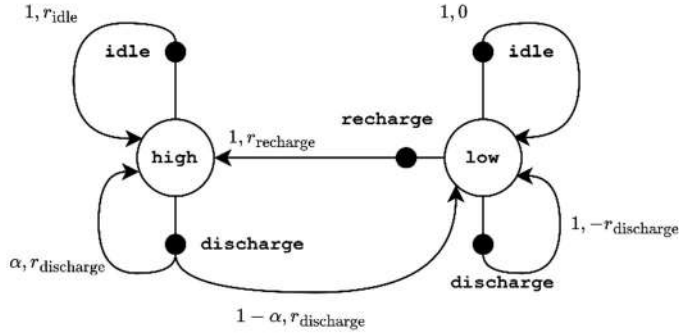
- verovatnoća da baterija ostane u stanju *high* iznosi α ;
- verovatnoća prelaza u stanje *low* iznosi $1 - \alpha$.

U oba slučaja dodeljuje se ista nagrada $r_{discharg}$. Primer prelaza sa nagradom 0 je

situacija u kojoj sistem, nakon primene akcije *idle* u stanju *low*, ostane u istom stanju. Pokušaj agenta da u stanju *low* primeni akcija *discharge*, pri čemu baterija ostaje u *low* stanju, obeshrabruje se negativnom nagradom.

Tabela 1. Primer MDP-a koji modeluje korišćenje električnog vozila: $p(s'|s, a)$ i $r(s, a, s')$ predstavljaju verovatnoću prelaska iz stanja s u stanje s' preduzimanjem akcije a , tim redosledom. Neki prelazi su nemogući, te imaju verovatnoću 0 i nemaju dodeljenu nagrade.

s	a	s'	$p(s' s, a)$	$r(s, a, s')$
high	idle	high	1	r_{idle}
high	idle	low	0	-
low	idle	low	1	0
low	idle	high	0	-
high	discharge	high	α	$r_{discharge}$
high	discharge	low	$1-\alpha$	$r_{discharge}$
low	discharge	low	1	$-r_{discharge}$
low	discharge	high	0	-
low	recharge	high	1	$r_{recharge}$
low	recharge	low	0	-



Slika 7. Graf prelaza u Markovljevom procesu odlučivanja

Cilj agenta je da nauči da odabere najpogodnije akcije u datim stanjima okruženja, odnosno, akcije koje vode ka maksimalnoj dugoročnoj nagradi, to jest, očekivanom dobitku (eng. *gain*):

$$G_t \doteq R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}, \quad (5)$$

gde:

- t predstavlja trenutni vremenski korak u kojem agent donosi odluku,
- R_{t+k+1} predstavlja neposrednu nagradu koju agent dobija u vremenskom koraku $t + k + 1$,
- γ predstavlja hiper-parametar poznat kao faktor umanjenja (eng. *discount rate*) za koga važi $0 \leq \gamma \leq 1$. Kada je $\gamma = 0$, agent maksimizira samo neposrednu

nagradu, dok se sa povećanjem γ prema vrednosti 1 uvećava fokus agenta na buduće nagrade.

Primer 2.2 prikazuje značaj preciznog postavljanja ciljeva prilikom definisanja funkcije nagrade za agenta. U njemu su prikazana tri scenarija koja predstavljaju različite pristupe nagrađivanju, u zavisnosti od cilja: korišćenja električnog vozila kao skladišta energije, za prevoz, ili za kombinaciju oba. Ključna poruka je da funkcija nagrade mora biti usklađena sa konkretnim ciljem koji agent treba da ostvari.

Primer 2.2 Vrednosti stanja i akcije iz primera 2.1 preslikavaju se u novi kontekst. Prilikom definisanje nagrade za električno vozilo (eng. *Electric Vehicle*, EV), ključno je jasno postaviti cilj koji agent treba da postigne. U ovom primeru posmatraju se tri različita cilja:

- (1) EV se koristi kao sistem za skladištenje električne energije,
- (2) EV se koristi isključivo za prevoz i
- (3) kombinacija prethodna dva cilja.

Okruganje predstavlja EV priključeno na kućni energetski sistem, pri čemu se posmatra period od jednog dana (od 00:00 do 23:59). Sistem je pojednostavljen tako što je unapred poznato kada je EV u upotrebi i kada nije priključeno na kuću.

U prvom scenariju, gde se EV koristi kao baterija, nagrade za prelaze stanja definišu se tako da podstiču akciju *recharge* kada je baterija u stanju *low*, dok se akcija *discharge* favorizuje kada je baterija u stanju *high*:

- Nagrade $r(\text{high}, \text{discharge}, \text{high})$, $r(\text{high}, \text{discharge}, \text{low})$ i $r(\text{low}, \text{recharge}, \text{high})$ iznose +1, označavajući pozitivnu nagradu,
- Nagrada $r(\text{low}, \text{discharge}, \text{low})$ iznosi -1, predstavljajući negativnu nagradu.

U drugom scenariju, uvodimo indikator lokacije koji pokazuje da li je EV priključen na kuću (*home*) ili nije (*away*). Pored toga, dodaje se nova akcija “vozi” (*drive*). Funkcija nagrade definisana je tako da podstiče vožnju EV kada je baterija u stanju *high* i povratak kući kada je baterija u stanju *low*. Sa druge strane, vožnja se obeshrabruje kada je baterija u stanju *low*. Radi pojednostavljenja, nagrade se dodeljuju samo za ključna stanja. Nagrade su definisane na sledeći način:

- Nagrade $r([\text{high}, \text{home}], \text{drive}, [\text{high}, \text{away}])$, $r([\text{high}, \text{away}], \text{drive}, [\text{high}, \text{away}])$, $r([\text{low}, \text{away}], \text{drive}, [\text{low}, \text{home}])$ iznose +1,
- Nagrade $r([\text{low}, \text{away}], \text{drive}, [\text{low}, \text{away}])$, $r([\text{low}, \text{home}], \text{drive}, [\text{low}, \text{away}])$ iznose -1.

U trećem scenariju, cilj je uspostavljanje ravnoteže između dva prethodno definisana cilja – korišćenja EV kao baterije dok je priključeno na kuću i efikasne upotrebe energije tokom vožnje. U stanje dodajemo faktor vremena kako bi se model prilagodio planiranju vožnje. Radi pojednostavljenja, vreme je diskretizovano, a periodi u kojima vozač planira polazak ili dolazak unapred su poznati. Vremenska komponenta uvedena u stanje se

računa kao apsolutna razlika između planiranog vremena polaska ili dolaska i trenutnog vremena. Ovaj scenario se neće detaljnije razmatrati zbog njegove složenosti.

Za opisan MDP, politika (eng. *policy*) π je funkcija koja svakom stanju s pridružuje raspodelu verovatnoća nad skupom akcija $A(s)$. Ako agent u trenutku t postupa prema politici π , oznaka $\pi(a|s)$ označava verovatnoću da će agent izvršiti akciju a u stanju s (Nikolic and Zecevic 2019). Deterministička politika je slučaj u kom agent izabira tačno jednu akciju u svakom stanju. Stohastička politika pridružuje verovatnoće akcijama, gde $\pi(s, a)$ predstavlja raspodelu verovatnoća nad skupom akcija za dato stanje. Pri izboru akcije, agent može birati akciju sa najvećom verovatnoćom ili koristiti različite strategije za balansiranje istraživanja i eksploatacije.

Funkcija vrednosti stanja (eng. *state value function*), poznata i kao vrednost stanja $v_\pi(s)$, definisana je kao očekivana dugoročna nagrada za primenu politike π počev od stanja s :

$$v_\pi(s) \doteq \mathbb{E}_\pi[G_t | S_t = s] = \mathbb{E}_\pi \left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \middle| S_t = s \right], \text{ za sve } s \in S \quad (6)$$

gde:

- G_t označava maksimalnu dugoročnu nagradu definisanu u jednačini (5)
- S_t označava stanje u trenutku t
- R_{t+k+1} označava neposrednu nagradu koju agent dobija u vremenskom koraku $t + k + 1$
- γ predstavlja faktor umanjenja.

Funkcija vrednosti akcije u određenom stanju (eng. *state-action value function*), poznata i kao vrednost akcije a u stanju s , definiše se kao očekivana dugoročna nagrada koju agent dobija kada u stanju s odabere akciju a , nakon čega nastavlja da postupa u skladu sa politikom π . Ova funkcija formalno je definisana na sledeći način:

$$q_\pi(s, a) \doteq \mathbb{E}_\pi[G_t | S_t = s, A_t = a] = \mathbb{E}_\pi \left[\sum_{k=0}^{\infty} \gamma^k R_{t+k+1} \middle| S_t = s, A_t = a \right], \quad (7)$$

gde:

- G_t označava maksimalnu dugoročnu nagradu definisanu u jednačini (5),
- S_t označava stanje u trenutku t ,
- A_t označava akciju agenta u stanju S_t ,
- R_{t+k+1} označava neposrednu nagradu koju agent dobija u vremenskom koraku $t + k + 1$,
- γ predstavlja faktor umanjenja.

Funkcija vrednosti stanja (v) koristi se za procenu dugoročne nagrade, pri čemu razmatra isključivo trenutno stanje. Sa druge strane, funkcija vrednosti akcije u određenom stanju (q), prilikom procene dugoročne nagrade uzima u obzir i stanje i akciju.

Obe funkcije vrednosti zadovoljavaju rekurentne relacijame (eng. *recursive relationship*) koje se nazivaju Belmanovim jednačinama (eng. *Bellman equation*) (Sutton and Barto 2014). Za svako stanje s i svaku politiku π važi Belmanova jednačina za v_π :

$$\begin{aligned}
v_\pi(s) &\doteq \mathbb{E}_\pi[G_t | S_t = s] \\
&= \mathbb{E}_\pi[R_{t+1} + \gamma G_{t+1} | S_t = s] \\
&= \sum_a \pi(a|s) \sum_{s',r} p(s',r|s,a) [r + \gamma \mathbb{E}_\pi[G_{t+1} | S_{t+1} = s']] \quad (8) \\
&= \sum_a \pi(a|s) \sum_{s',r} p(s',r|s,a) [r + \gamma v_\pi(s')].
\end{aligned}$$

Slično, za svako stanje s i svaku politiku π važi Belmanova jednačina za q_π :

$$\begin{aligned}
q_\pi(s, a) &\doteq \mathbb{E}_\pi[G_t | S_t = s, A_t = a] \\
&= \mathbb{E}_\pi[R_{t+1} + \gamma G_{t+1} | S_t = s, A_t = a] \\
&= \sum_{s',r} p(s',r|s,a) [r + \gamma \mathbb{E}_\pi[G_{t+1} | S_{t+1} = s']] \quad (9) \\
&= \sum_{s',r} p(s',r|s,a) [r + \gamma v_\pi(s')] \\
&= \sum_{s',r} p(s',r|s,a) [r + \gamma \sum_{a'} \pi(a'|s') q_\pi(s', a')].
\end{aligned}$$

Funkcije vrednosti v_π , odnosno q_π predstavljaju jedinstveno rešenje odgovarajućih Belmanovih jednačina. Ove jednačine predstavljaju osnovu mnogih pristupa za računanje, aproksimaciju i učenje v_π i q_π , omogućavajući evaluaciju politika. Rešavanje problema učenja potkrepljivanjem svodi se na pronalaženje optimalne politike π_* , prema kojoj agent donosi odluke u okruženju s ciljem maksimizacije očekivane dugoročne nagrade. Politiku π smatramo boljom ili jednakom politici π' ($\pi \geq \pi'$) ako za sva stanja $s \in S$ važi $v_\pi(s) \geq v_{\pi'}(s)$. Uvek postoji najmanje jedna optimalna politika π_* koja je bolja ili jednaka bilo kojoj drugoj politici (Nikolic and Zecevic 2019).

Iako može postojati više optimalnih politika, sve dele iste funkcije v i q :

$$\begin{aligned}
v_*(s) &\doteq \max_{\pi} v_\pi(s), \\
q_*(s, a) &\doteq \max_{\pi} q_\pi(s, a), \quad (10)
\end{aligned}$$

za svako stanje $s \in S$ i akciju $a \in A$. Funkcije vrednosti v_* i q_* su optimalne funkcije vrednosti. Takođe, v_* se može definisati preko q_* na sledeći način:

$$v_*(s) \doteq \max_a q_*(s, a), \quad (11)$$

Funkcije v_* i q_* mogu da se izraze Belmanovim jednakostima na sledeći način:

$$\begin{aligned}
v_*(s) &= \max_a \sum_{s',r} p(s',r|s,a) [r + \gamma v_*(s')], \\
q_*(s, a) &= \sum_{s',r} p(s',r|s,a) \left[r + \gamma \max_{a'} q_*(s', a') \right]. \quad (12)
\end{aligned}$$

Poznajući funkcije v_* ili q_* moguće je izračunati barem jednu optimalnu politiku. Najpoželjnija akcija a u stanju s je ona koja daje najveću očekivanu nagradu, gde je optimalna deterministička politika definisana na sledeći način:

$$\pi_*(a|s) = \begin{cases} 1 & \text{ako važi } a = \underset{a}{\operatorname{argmax}} q_*(s, a), \\ 0 & \text{inače.} \end{cases}$$

Posmatrajući q_* , najveća očekivana dugoročna nagrada pri preuzimanju akcije a u stanju s zavisi od neposredne dobijene nagrade r i najveće dugoročne nagrade iz novog stanja s' u koje agent prelazi nakon preduzimanja akcije a' . Težinski faktor u ovom proračunu predstavlja verovatnoća prelaska iz stanja s u stanje s' pri akciji a , koja rezultuje nagradom r , data u jednačini (4).

Osnovni problem učenja potkrepljivanjem je određivanje optimalnih funkcija vrednosti za stanja i akcije u stanju, odnosno, pronalaženje optimalne politike. U slučaju poznatog okruženja, agent poseduje kompletan i tačan model dinamike okruženja. Ako je okruženje MDP, taj model je definisan funkcijom prelaska stanja (jednačina (4)). U tom slučaju, funkcije vrednosti mogu se tačno izračunati korišćenjem Belmanovih jednačina. Međutim, u realnim scenarijima, okruženje je često nepoznato, što znači da agent nema pristup kompletnom i tačnom modelu okruženja. U takvim situacijama pribegava se metodama približnog računanja ili učenju kroz interakciju sa okruženjem.

U slučaju nepoznatog okruženja, agent primenjuje različite akcije kako bi istražio okruženje sa ciljem da nauči optimalnu politiku π_* . U odgovarajućem MDP, agent sakuplja informaciju o funkciji (4) na dva načina :

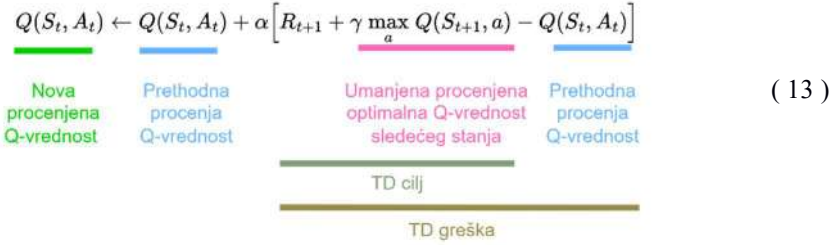
- *U skladu sa politikom* (eng. *on-policy*) – agent optimizuje istu politiku koju koristi za odabir akcija. Drugim rečima, politika koju agent koristi za interakciju sa okruženjem je ista politika koju pokušava da poboljša.
- *Mimo politike* (eng. *off-policy*) – agent koristi različite politike za odabir akcija (istraživanje) od onih koje optimizuje (ciljne politike). To znači da agent može da koristi jednu politiku za istraživanje okruženja, dok istovremeno uči i poboljšava drugu politiku.

Potpoglavlja 2.2.1.1 i 2.2.1.2 obrađuju Q-učenje (eng. *Q-Learning*), duboku Q-mrežu (eng. *Deep Q-Network*, DQN) i dvostruku duboku Q-mrežu (eng. *Double Deep Q-Network*, DDQN), koji pripadaju algoritmima učenja potkrepljivanjem mimo politike. Q-učenje predstavlja klasičan algoritam koji koristi tabelarni pristup za procenu očekivane dugoročne nagrade. DQN proširuje ovaj pristup zamenom tabele neuronskom mrežom, što omogućava primenu u okruženjima sa velikim ili kontinuiranim skupovima stanja. DDQN dodatno unapređuje DQN uvođenjem dve odvojene mreže – jedne za selekciju akcije, a druge za njenu evaluaciju – čime se postiže stabilnije i preciznije učenje. Potpoglavlje 2.2.1.3 obrađuje algoritam proksimalne optimizacije politike, koji pripada grupi algoritama učenja potkrepljivanjem u skladu sa politikom.

2.2.1.1 Q-učenje

Metoda učenja trenutnih razlika (eng. *temporal difference*, TD) predstavlja pristup učenju na osnovu iskustva, bez potrebe za eksplicitnim modelom okruženja. TD metode vrše procene funkcije vrednosti koristeći razliku između trenutne procene funkcije vrednosti i novodobijene informacije dobijene iz narednog koraka.

Jedan od najpoznatijih TD algoritma je Q-učenje (Sutton and Barto 2014). Kod Q-učenja, procena funkcije vrednosti akcije u datom stanju (Q-funkcija) se ažurira na osnovu neposredne nagrade i maksimalne procenjene vrednosti narednog stanja, pri svakoj interakciji sa okruženjem, bez potrebe za eksplicitnim modelom dinamike okruženja (Lonza 2019). Ovaj algoritam zasniva se na sledećem pravilu ažuriranja funkcije vrednosti (Watkins 1989):

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha \left[R_{t+1} + \gamma \max_a Q(S_{t+1}, a) - Q(S_t, A_t) \right] \quad (13)$$


gde α označava stopu učenja (eng. *learning rate*), a γ faktor umanjavanja. TD cilj predstavlja zbir nagrade R_{t+1} nakon preduzimanja akcije A_t i maksimalne Q-vrednosti sledećeg stanja S_{t+1} , procenjene na osnovu pohlepne politike. TD greška se računa kao razlika TD cilja i prethodne procenjene Q-vrednosti.

Kao politika ponašanja najčešće se koristi ϵ -pohlepna politika (eng. *ϵ -greedy policy*), koja predstavlja kompromis između istraživanja i eksploatacije (eng. *exploration-exploitation tradeoff*). Većinu vremena agent se oslanja na eksploataciju – sa verovatnoćom $1 - \epsilon$ bira akciju koja maksimizuje trenutnu procenu Q-funkcije, odnosno, akciju za koju se očekuje najveća dugoročna nagrada. Da bi obezbedio istraživanje, odnosno prikupljanje novih informacija o okruženju, agent sa malom verovatnoćom $\epsilon > 0$ nasumično bira jednu od dostupnih akcija, pri čemu sve imaju jednaku verovatnoću odabira. Odabir akcije prema ϵ -pohlepnoj politici može se formalno izraziti na sledeći način:

$$\pi(s, a) = \begin{cases} \frac{\epsilon}{|A(s)|}, & \text{za sve akcije } a \neq \underset{a'}{\operatorname{argmax}} Q(s, a'), \\ 1 - \epsilon + \frac{\epsilon}{|A(s)|}, & \text{za akciju } a = \underset{a'}{\operatorname{argmax}} Q(s, a'). \end{cases} \quad (14)$$

Algoritam 1 prikazuje pseudokod za algoritam Q-učenja. Algoritam započinje inicijalizacijom vrednosti Q-funkcije na proizvoljne vrednosti za sve parove stanja i akcija. Vrednosti terminalnih stanja se inicijalizuju na 0, dok se ostale vrednosti $Q(s, a)$ tipično postavljaju na 0 radi pojednostavljenja, iako je moguće koristiti i druge početne vrednosti. Tokom svake epizode, u svakom vremenskom koraku t , bira se akcija A_t u skladu sa politikom ponašanja, koja je najčešće ϵ -pohlepna politika. Nakon što se akcija A_t prosledi okruženju, agent dobija sledeće stanje S_{t+1} , nagradu R_{t+1} i indikator kraja epizode *done*. Na osnovu ovih informacija agent ažurira procenu Q-vrednosti prema pravilu ažuriranja iz jednačine (13). Na kraju svake epizode, stopa istraživanja ϵ se obično smanjuje prema unapred definisanom rasporedu, kako bi agent postepeno prelazio sa istraživanja na eksploataciju.

Po završetku procesa učenja, agent definiše ciljnu politiku $\pi^*(s)$ kao determinističku politiku koja u svakom stanju bira akciju sa najvećom naučenom Q-vrednošću:

$$\pi^*(s) = \underset{a}{\operatorname{argmax}} Q(s, a).$$

Na ovaj način, agent prelazi sa istraživanja na eksploataciju i koristi naučene vrednosti za optimalno donošenje odluka.

Algoritam 1. Q-učenje

```

Ulaz:
• broj_epizoda – ukupan broj epizoda
•  $\alpha \in (0,1]$  – stopa učenja
•  $\gamma$  – faktor umanjenja
•  $\varepsilon$  – početna stopa istraživanja
•  $\varepsilon_{min}$  – minimalna vrednost  $\varepsilon$ 
•  $\varepsilon_{decay}$  – stopa smanjenja  $\varepsilon$ 
•  $Q(s, a)$  – inicijalizovana Q-tabela (na primer,  $Q(s, a) = 0$  za sve  $s \in S$  i  $a \in A(s)$ ).
1 for epizoda in broj_epizoda do
2   Restartuj okruženje i inicijalizuj početno stanje  $S_0$ 
3   for  $t = 1$  do kraja epizode do (dok  $done = False$ )
4     Izaberi akciju  $A_t$  koristeći politiku ponašanja (na primer,  $\varepsilon$ -pohlepnu politiku prema
       jednačini (13))
5     Izvrši akciju  $A_t$  i dobavi  $(S_{t+1}, R_{t+1}, done)$  iz okruženja
6     Ažuriraj Q-vrednost prema pravilu ažuriranja iz jednačine (13)
7     Postavi  $S_t \leftarrow S_{t+1}$ 
8   end
9   Ažuriraj stopu istraživanja:  $\varepsilon_{t+1} \leftarrow \max(\varepsilon_{min}, \varepsilon_t \cdot \varepsilon_{decay})$ 
10 end

```

2.2.1.2 Duboka Q-mreža i dvostruka duboka Q-mreža

Dimenzije Q-tabele eksponencijalno rastu sa porastom veličine skupova stanja i akcija, što dovodi do visokih zahteva za memorijom i otežava primenu klasičnog Q-učenja u složenim okruženjima. DQN algoritam (Mnih et al. 2013) razvijen je kako bi omogućio primenu Q-učenja u realističnim scenarijima sa velikim ili kontinuiranim prostorima stanja. Umesto eksplicitnog čuvanja Q-vrednosti u tabeli, DQN koristi neuronsku mrežu za aproksimaciju funkcije vrednosti akcije (Q-funkcije) na osnovu karakteristika stanja. Neuronska mreža se obučava da precizno aproksimira optimalnu funkciju vrednosti $q_*(s, a)$. Na taj način, omogućena je generalizacija na stanja koja nisu eksplicitno posećena tokom učenja, kao i značajno efikasnije korišćenje memorijskih resursa.

U DQN algoritmu koristi se mehanizam poznat kao memorija za ponovno iskustvo (*experience replay*), koji omogućava efikasnije i stabilnije učenje. U svakom vremenskom koraku, iskustvo agenta u obliku četvorke (s_t, a_t, r_t, s_{t+1}) se smešta u memoriju D . Ova memorija akumulira podatke iz više epizoda, omogućavajući višestruko korišćenje prethodnih iskustva tokom treniranja mreže. Veličina memorije predstavlja jedan od hiperparametara algoritma. Za ažuriranje parametara mreže, iz memorije D se nasumično uzorkuje podskup četvorki, čime se prekida korelacija između uzastopnih uzoraka. Nasuprot tome, u klasičnom Q-učenju ažuriranja se vrše direktno na osnovu uzastopnih stanja, što može izazvati visoku varijansu ažuriranja usled međuzavisnosti uzoraka.

U DQN algoritmu koriste se dve neuronske mreže: mreža politike (Q-mreža) i ciljna mreža (eng. *target network*) (Mnih et al. 2015). Ulaz u Q-mrežu čine karakteristike stanja s_t , dok izlaz predstavlja vektor procenjenih Q-vrednosti za sve dozvoljene akcije. Ciljna mreža ima istu arhitekturu kao mreža politike, ali se njeni parametri θ' ažuriraju ređe – periodičnim kopiranjem vrednosti iz Q-mreže (parametara θ). Ovaj mehanizam, poznat kao zamrzavanje ciljne mreže (eng. *target freezing*), doprinosi stabilnijem učenju tako što sprečava oscilacije u ciljnim vrednostima. Parametri Q-mreže ažuriraju se minimizacijom greške između procenjene Q-vrednosti $Q(s_{t+1}, a_t; \theta)$ i ciljne vrednosti y , definisane kao:

$$y = \begin{cases} r_{t+1}, & \text{ako je } s_{t+1} \text{ terminalno stanje,} \\ r_{t+1} + \gamma \max_{a'} Q(s_{t+1}, a'; \theta') & \text{inače.} \end{cases} \quad (15)$$

Greška koja se minimizuje tokom obuke definiše se sledećom funkcijom gubitka:

$$L(\theta) = \mathbb{E}_{(s,a,r,s') \sim \mathcal{D}} \left[(y - Q(s, a; \theta))^2 \right]. \quad (16)$$

Algoritam 2 prikazuje tok DQN algoritma. Obučavanje se odvija u dve faze:

- **Uzorkovanje:** agent izvršava akcije u okruženju i beleži iskustva u obliku torki (s_t, a_t, r_t, s_{t+1}) , koje se smeštaju u memoriju za ponovno iskustvo.
- **Obučavanje:** nasumično se uzorkuje podskup torki iz memorije, na osnovu kojih se ažuriraju parametri mreže.

Prvi korak podrazumeva inicijalizaciju parametara, uključujući kapacitet memorije za ponovno iskustvo $\mathcal{D}$, kao i početne vrednosti parametara mreže politike θ i ciljne mreže $\theta' = \theta$. Na početku svake epizode okruženje se resetuje i dobija se početno stanje. U svakom vremenskom koraku agent bira akciju korišćenjem ε -pohlepne politike zasnovane na procenama Q-vrednosti iz mreže politike. Nakon primene akcije nad okruženjem, dobijaju se sledeće stanje, nagrada i indikator kraja epizode. Iskustvo u obliku torke $(s_t, a_t, r_t, s_{t+1}, done)$ smešta se u memoriju za ponovno iskustvo. Kada u memoriji postoji dovoljan broj torki, nasumično se uzorkuje podskup, na osnovu koga se izračunava ciljna vrednost y prema jednačini (15). Mreža politike se potom obučava minimizacijom greške između $Q(s_{t+1}, a_t; \theta)$ i y , kako je definisano formulom (16). Konačno, parametri ciljne mreže θ' se periodično ažuriraju kopiranjem iz mreže politike.

Algoritam 2. DQN algoritam

Ulaz:

- *broj_epizoda* – ukupan broj epizoda
- *N_{batch}* – veličina podskupa torki uzorkovanih iz memorije za ponovno iskustvo
- *D* – kapacitet memorije za ponovno iskustvo
- *C* – broj koraka nakon kojih se parametri mreže politike kopiraju u ciljnu mrežu

```

1 Postavka kapaciteta memorije za ponovljeno iskustvo na D
2 Inicijalizacija parametra mreže politike  $\theta$  i postavka  $\theta' \leftarrow \theta$ 
3 for epizoda=1 do broj_epizoda:
4     Restartuj okruženje i postavi početno stanje  $S_0$ 
5     for t = 1 do kraja epizode do: (dok done = False)
6         Izabrati akciju  $A_t$  korišćenjem politike ponašanja (npr.,  $\varepsilon$ -pohlepne politike (13))
7         Izvršiti akciju  $A_t$  i dobiti  $(S_{t+1}, R_t, done)$  iz okruženja
8         Sместiti torku  $(S_t, A_t, R_t, S_{t+1}, done)$  u memoriju za ponovno iskustvo
9         Uzorkovati nasumični podskup od  $N_{batch}$  torki iz memorije za ponovno iskustvo
```

```

10      |   | Izračunati ciljnu vrednost y korišćenjem jednačine ( 15 )
11      |   | Ažurirati parametre  $\theta$  minimizacijom gubitka iz jednačine ( 16 )
12      |   | Ako je broj koraka deljiv sa C, postavi  $\theta' \leftarrow \theta$ 
13      | end
14 end

```

Jedan od ključnih problema DQN algoritma je tendencija ka precenjivanju vrednosti akcija, jer se ista mreža koristi i za izbor akcije i za procenu njene vrednosti. Algoritam DDQN unapređuje DQN razdvajanjem ovih uloga na dve odvojene neuronske mreže:

- Mreža politike (eng. *policy network*, Q) koristi se za izbor akcije sa najvećom procenjenom Q-vrednošću u sledećem stanju.
- Ciljna mreža Q' (eng. *target network*) koristi se za procenu vrednosti izabrane akcije u sledećem stanju.

Ova modifikacija redukuje pristrasnost ka precenjivanju Q-vrednosti i doprinosi stabilnijem učenju (Hasselt et al. 2015). Pravilo ažuriranja u DDQN algoritmu ima sledeći oblik:

$$Q(S_t, A_t) \leftarrow Q(S_t, A_t) + \alpha \left[R_{t+1} + \gamma Q' \left(S_{t+1}, \underset{a}{\operatorname{argmax}} Q(S_{t+1}, a) \right) - Q(S_t, A_t) \right]. \quad (17)$$

Ciljna vrednost y_t koja se koristi za treniranje mreže politike, računa se prema sledećem izrazu: $a_{\max} = \underset{a'}{\operatorname{argmax}} Q(s_{t+1}, a'; \theta)$ (18)

$$y_t = \begin{cases} r_{t+1}, & \text{ako je } s_{t+1} \text{ terminalno stanje,} \\ r_{t+1} + \gamma Q'(s_{t+1}, a_{\max}; \theta'), & \text{inače.} \end{cases}$$

Zabeležene torke iz memorije za ponovno iskustvo koriste se za ažuriranje mreže politike Q , dok se parametri ciljne mreže Q' ažuriraju periodično, kopiranjem iz mreže politike. U poređenju sa Algoritam 2, ključna razlika u pseudokodu za DDQN nalazi se u liniji 10, gde se ciljna vrednosti računa prema formuli (18), umesto korišćenja maksimalne Q-vrednosti iz iste mreže.

2.2.1.3 Proksimalna optimizacija politike

Za razliku od Q-učenja, DQN i DDQN algoritama, koji pripadaju metodama učenja sa funkcijom vrednosti, algoritam PPO (eng. *Proximal Policy Optimization*) pripada grupi metoda gradijenta politike (eng. *policy gradient methods*). Ove metode uče politiku $\pi(a|s)$ direktno, bez eksplicitnog modelovanja Q-funkcije. Pregled ključnih razlika između Q-učenja, DQN/DDQN i PPO algoritama dat je u tabeli 2.

Tabela 2. Uporedna analiza algoritama Q-učenje, DQN/DDQN i PPO

Karakteristike	Q-učenje	DQN/DDQN	PPO
Tip metode	Tabelarna metoda sa funkcijom vrednosti	Funkcija vrednosti sa aproksimacijom pomoću neuronske mreže	Gradijent politike (akter-kritičar pristup)
Reprezentacija	Q-tabela	Neuronska mreža aproksimira $Q(s, a)$	Politika $\pi(a s)$
Funkcija vrednosti	$Q(s, a)$	$Q(s, a)$	$V(s), A(s, a)$

Politika ponašanja	ε - pohlepna politika	ε -pohlepna politika	Direktna politika sa ograničenjem promene (<i>clipping</i>)
Ažuriranje parametara	TD pravilo	MSE između Q-vrednosti i ciljne vrednosti	Maksimizacija ciljne funkcije L^{CLIP}
Stabilizacija učenja	Nema	Memorija za ponovno iskustvo i ciljna mreža	Odsecanje promene politike, funkcija prednosti (<i>advantage</i>), podskupovi (<i>minibatch</i>)
Prostor akcija	Diskretan	Diskretan	Podržava diskretan i kontinuiran prostor akcija
Prostor stanja	Diskretan i niskodimenzioni	Podržani su kontinuirani i visokodimenzioni prostori stanja	Podržani su kontinuirani i visokodimenzioni prostori stanja

PPO algoritam se zasniva na akter-kritičar (eng. *actor-critic*) metodama učenja, koje pripadaju kategoriji učenja u skladu sa politikom. Akter (eng. *actor*) predstavlja neuronsku mrežu koja uči politiku $\pi(a|s)$, odnosno, verovatnoću izbora akcije u datom stanju. Kritičar (eng. *critic*) je neuronska mreža koja procenjuje kvalitet ponašanja aktera, najčešće kroz procenu funkcije vrednosti stanja $V(s)$. Na osnovu te procene, izračunava se funkcija prednosti $A(s, a)$, koja se koristi za ažuriranje parametara aktera. Time se unapređuje politika tako da povećava verovatnoću izbora akcija koje vode ka većoj dugoročnoj nagradi.

Proces učenja u PPO algoritmu podrazumeva procenu ciljne funkcije vrednosti V_t^{target} i funkcije prednosti A_t . Ove vrednosti izračunavaju se pre ažuriranja parametara mreža, na osnovu prethodne verzije politike $\pi_{\theta_{old}}$ funkcije vrednosti stanja $V_{\theta_{old}}$ (Schulman et al. 2017). Ciljna vrednost V_t^{target} predstavlja umanjenju dugoročnu nagradu dobijenu nakon izvršene akcije a_t u stanju s_t , i računa se kao (Mnih et al. 2016):

$$V_t^{target} = r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \dots + \gamma^{n-1} r_{t+n-1} + \gamma^n V_{\theta_{old}}(s_{t+n}). \quad (19)$$

U (19), parametar n označava broj koraka unapred tokom kojih se akumuliraju nagrade pre nego što se izvrši procena vrednosti iz stanja s_{t+n} korišćenjem aproksimatora vrednosti (eng. *bootstrap*) (Nikolic and Zecevic 2019).

Funkcija prednosti A_t meri koliko je konkretan ishod – odnosno, dugoročna nagrada – bio bolji ili lošiji od očekivane vrednosti stanja, koju procenjuje kritičar. Formalno, definiše se kao razlika između ciljne vrednosti i vrednosti trenutnog stanja (Schulman et al. 2017), (Mnih et al. 2016):

$$A_t = V_t^{target} - V_{\theta_{old}}(s_t). \quad (20)$$

Empirijski je pokazano da akter-kritičar metode često pate od nestabilnosti, pri čemu čak i jedan loše odabran korak ažuriranja može značajno narušiti performanse (Sun et al. 2025). PPO algoritam ublažava ovu nestabilnost ograničavanjem veličine promene politike putem mehanizma odsecanja (eng. *clipping*) u ciljnoj funkciji optimizacije. Ciljna funkcija optimizacije L^{CLIP} koristi se za ažuriranje mreže aktera (eng. *PPO policy network*) (Bick, 2021) i definisana je kao:

$$L^{CLIP}(\theta) = \mathbb{E}_t[\min(p_t(\theta)A_t, \text{clip}(p_t(\theta), 1 - \epsilon, 1 + \epsilon)A_t)], \quad (21)$$

gde je $p_t(\theta)$ odnos između trenutne i prethodne politike (Schulman et al. 2017) (Kakade and Langford 2002), definisan kao:

$$p_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)}. \quad (22)$$

Odnos $p_t(\theta)$ predstavlja količnik verovatnoća: ako je $p_t(\theta) > 1$, akcija a_t u stanju s_t ima veću verovatnoću odabira pod trenutnom politikom π_θ nego pod prethodnom politikom. Suprotno važi za dok za $0 < p_t(\theta) < 1$.

Ciljna funkcija L^{CLIP} ima dva ključna efekta:

1. Konzervativnu procenu učinka politike – vrednost funkcije uzima minimum između standardnog (eng. *unclipped*) izraza $p_t(\theta)A_t$ i ograničenog (eng. *clipped*) izraza $clip(p_t(\theta), 1 - \epsilon, 1 + \epsilon)A_t$. Time se ublažava uticaj prevelikih oscilacija u ažuriranju.
2. Sprečavanje prevelikih promena politike – funkcija onemogućava da se značajno poveća verovatnoća odabira akcije a_t u stanju s_t , naročito u slučajevima kada su prethodna ažuriranja već dovela do poboljšanja. Ova restrikcija se sprovodi putem faktora $clip(p_t(\theta), 1 - \epsilon, 1 + \epsilon)$, koji ograničava koliko se odnos verovatnoća između trenutne i prethodne politike može udaljiti od 1 (Bick, 2021):

$$clip(p_t(\theta), 1 - \epsilon, 1 + \epsilon) = \begin{cases} 1 - \epsilon, & \text{ako } p_t(\theta) < 1 - \epsilon, \\ 1 + \epsilon, & \text{ako } p_t(\theta) > 1 + \epsilon, \\ p_t(\theta), & \text{inače.} \end{cases} \quad (23)$$

Ovaj efekat ilustrovan je na slici 8, gde se prikazuje kako se vrednost odnosa verovatnoća $p_t(\theta)$ održava unutar intervala $[1 - \epsilon, 1 + \epsilon]$, čak i nakon više epoha ažuriranja težina nad istim podacima za obuku.

Radi boljeg razumevanja načina na koji mehanizam odsecanja u PPO funkcioniše u praksi, tabela 3 prikazuje sve moguće kombinacije odnosa verovatnoća $p_t(\theta)$ i znaka funkcije prednosti A_t . Tabela prikazuje kada dolazi do odsecanja ciljne funkcije $L^{CLIP}(\theta)$, da li se parametri mreže ažuriraju, kao i u kom pravcu se politika menja:

- Kada je $1 - \epsilon \leq p_t(\theta) \leq 1 + \epsilon$ ciljna funkcija se ne odseca, nezavisno od znaka funkcije prednosti A_t . U redu 1, kada je $A_t > 0$, akcija je bolja od očekivane vrednosti svih dostupnih akcija u stanju s_t , te se verovatnoća njenog izbora povećava. U redu 2, kada je $A_t < 0$, akcija je lošija od očekivanja, pa se verovatnoća njenog izbora smanjuje. U oba slučaja, gradijent je prisutan i parametri politike se ažuriraju.
- U redovima 3 i 4, analiziraju se slučajevi kada je $p_t(\theta) < 1 - \epsilon$, što znači da trenutna politika dodeljuje akciji manju verovatnoću u poređenju sa prethodnom politikom. U redu 3, $A_t > 0$, što implicira da bi verovatnoću izbora akcije trebalo povećati. Funkcija se ne odseca i vrši se ažuriranje. U redu 4, $A_t < 0$, akcija je nepovoljna, a politika je već umanjila njenu verovatnoću. Funkcija se odseca, gradijent je jednak nuli i ne dolazi do ažuriranja težina.
- U redovima 5 i 6, slučajevi sa $p_t(\theta) > 1 + \epsilon$ ukazuju da trenutna politika značajno favorizuje akciji u odnosu na prethodnu politiku. U redu 5, $A_t > 0$, što znači da je akcija dobra, ali je politika već povećala verovatnoću više nego što

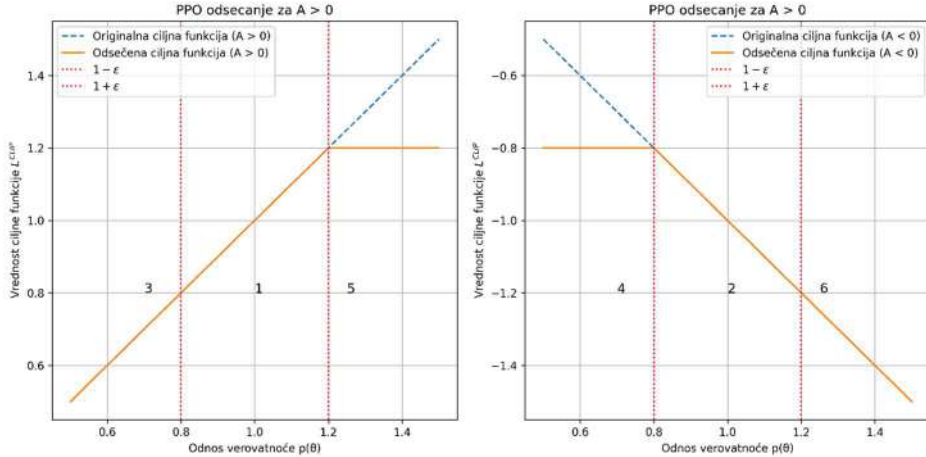
je dozvoljeno. Funkcija se odseca, gradijent je nula i nema ažuriranja težina. U redu 6, $A_t < 0$, što implicira da akciju treba obeshrabrili. Funkcija se ne odseca, gradijent je prisutan i verovatnoća izbora akcije se smanjuje.

Ukratko, politika se ažurira ukoliko:

- $p_t(\theta) \in [1 - \epsilon, 1 + \epsilon]$ (redovi 1 i 2),
- Kada $p_t(\theta) \notin [1 - \epsilon, 1 + \epsilon]$, ali znak A_t sugerise korektivnu promenu ka intervalu:
 - $p_t(\theta) < 1 - \epsilon, A_t > 0$ (red 3),
 - $p_t(\theta) > 1 + \epsilon, A_t < 0$ (red 6).

Tabela 3. Klasifikacija ponašanja ciljne funkcije $L^{CLIP}(\theta)$ i ažuriranje parametara u zavisnosti od odnosa verovatnoća $p_t(\theta)$ i funkcije prednosti A_t .

R e d	$p_t(\theta) > 0$	A_t	Povratna vrednost od min	Da li je ciljna funkcija $L^{CLIP}(\theta)$ odsečena	Znak A_t	Gradijent
1	$p_t(\theta) \in [1 - \epsilon, 1 + \epsilon]$	+	$p_t(\theta)A_t$	ne	+	✓
2	$p_t(\theta) \in [1 - \epsilon, 1 + \epsilon]$	−	$p_t(\theta)A_t$	ne	−	✓
3	$p_t(\theta) < 1 - \epsilon$	+	$p_t(\theta)A_t$	ne	+	✓
4	$p_t(\theta) < 1 - \epsilon$	−	$(1 - \epsilon)A_t$	da	−	0
5	$p_t(\theta) > 1 + \epsilon$	+	$(1 + \epsilon)A_t$	da	+	0
6	$p_t(\theta) > 1 + \epsilon$	−	$p_t(\theta)A_t$	ne	−	✓



Slika 8. Vizualizacija mehanizma odsecanja u PPO algoritmu za pozitivne i negativne vrednosti funkcije prednosti A_t . Leva strana prikazuje slučaj kada je $A_t > 0$: odsecanje ograničava rast ciljne funkcije za visoke vrednosti odnosa verovatnoća $p_t(\theta)$, koje označavaju da trenutna politika favorizuje akciju a_t znatno više nego prethodna. Desna strana prikazuje slučaj kada je $A_t < 0$: ograničava se pad vrednosti funkcije za niske vrednosti $p_t(\theta)$, koje ukazuju da trenutna politika daje znatno manju verovatnoću toj akciji u poređenju sa prethodnom politikom.

Algoritam 3 prikazuje pseudokod PPO algoritma, koji podrazumeva iterativno ažuriranje stohastičke politike π_θ (parametri mreže aktera θ) i funkcije vrednosti stanja $V_\phi(s)$ (parametri mreže kritičara ϕ). U svakoj iteraciji, agent izvršava trenutnu politiku π_{θ_k} u

okruženju i generiše skup trajektorija $D_k = \{\tau_i\}$. Na osnovu ovih trajektorija izračunavaju se umanjene dugoročne nagrade $\hat{R}_t$ i procene funkcija prednosti $\hat{A}_t$. Ove vrednosti se koriste za ažuriranje:

- Politike, maksimizacijom ciljne funkcije L^{CLIP} .
- Kritičara, minimizacijom srednje kvadratne greške između procene $V_\phi(s)$ i ciljne vrednosti $\hat{R}_t$.

Algoritam 3. PPO

Ulaz:

- ε – parameter odsecanja (eng. *clipping*)
- T – broj vremenskih koraka po trajektoriji

```

1  Inicijalizacija parametara politike  $\theta_0$  i vrednosne funkcije  $\phi_0$ 
2  for  $k = 1, 2, 3 \dots$  do
3      Izvrši politiku  $\pi_{\theta_k}$  u okruženju i generiši skup trajektorija  $D_k = \{\tau_i\}$ 
4      Izračunaj umanjene dugoročne nagrade  $\hat{R}_t$  za svaku trajektoriju
5      Izračunaj procenu prednosti  $\hat{A}_t$  (19) i (20) korišćenjem trenutne funkcije vrednosti  $V_{\phi_k}$ 
      Ažuriraj parametare aktera maksimizacijom ciljne funkcije  $L^{CLIP}$ :
6      
$$\theta_{k+1} = \underset{\phi}{\operatorname{argmax}} \frac{1}{|D_k|T} \sum_{\tau \in D_k} \sum_{t=0}^T \min \left( \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_k}(a_t|s_t)} A^{\pi_{\theta_k}}(s_t, a_t), g(\varepsilon, A^{\pi_{\theta_k}}(s_t, a_t)) \right)$$

      gde funkcija  $g$  primenjuje ograničenje (clipping) nad vrednostima prednosti.
      Ažuriraj parametre funkcije vrednosti minimizacijom srednje kvadratne greške:
7      
$$\phi_{k+1} = \underset{\phi}{\operatorname{argmin}} \frac{1}{|D_k|T} \sum_{\tau \in D_k} \sum_{t=0}^T (V_\phi(s_t) - \hat{R}_t)^2$$

8  end
```

2.2.1.4 Određivanje uticaja karakteristika stanja na odluke modela učenja potkrepljivanjem

Analiza uticaja karakteristika stanja na odluke modela učenja potkrepljivanjem zasniva se na kombinatornoj teoriji igara (eng. *Game Theory*) (Osborne and Rubinstein 2006), a posebno na konceptu *Shapley* vrednosti (Shapley 1953). Cilj je da se kvantifikuje uticaj svake karakteristike stanja na akciju koju model bira, čime se dobija interpretacija ponašanja modela učenja potprepljivanjem.

U kooperativnoj igri, koaliciona igra (eng. *coalition game*) definiše se nad skupom od N igrača, gde svaki igrač doprinosi ukupnoj vrednosti koalicije (Osborne and Rubinstein 2006). U kontekstu mašinskog učenja, igrači odgovaraju pojedinačnim karakteristikama stanja, dok vrednost koalicije predstavlja izlaz modela.

Značaj karakteristike i određuje se pomoću *Shapley* vrednosti, koja se za karakteristiku i računa prema sledećoj formuli (Osborne and Rubinstein 2006):

$$\phi_i = \frac{1}{|N|!} \sum_{S \subseteq N \setminus \{i\}} |S|! (|N| - |S| - 1)! [f(S \cup \{i\}) - f(S)]$$

gde:

- N označava skup svih karakteristika u modelu,
- $S \subseteq N \setminus \{i\}$ predstavlja proizvoljan podskup karakteristika koji ne sadrži i ,
- $|S|! (|N| - |S| - 1)!$ predstavlja broj permutacija skupa karakteristika u kojima se podskup S javlja kao skup koji prethodi karakteristici i ,

- $f(S)$ predstavlja vrednost koalicije za skup karakteristika S ,
- $f(S \cup \{i\})$ predstavlja vrednost koalicije kada se karakteristika i doda u skup S .

U ovoj formuli izraz $f(S \cup \{i\}) - f(S)$ predstavlja marginalni doprinos karakteristike i koaliciji S , odnosno promenu vrednosti koalicije koja nastaje dodavanjem karakteristike i u skup S . Tumačenje znaka marginalnog doprinosa je sledeće:

- Pozitivna vrednost znači da karakteristika i povećava vrednost koalicije.
- Negativna vrednost znači da karakteristika i smanjuje vrednost koalicije.
- Vrednost bliska nuli ukazuje da dodavanje karakteristike i nema značajan uticaj na koaliciju.

SHAP vrednost karakteristike i predstavlja primenu *Shapley* vrednosti u interpretabilnosti modela mašinskog učenja i dobija se kao prosečan marginalni doprinos karakteristike i izračunat preko svih mogućih podskupova S . Na taj način SHAP vrednost kvantifikuje prosečan efekat karakteristike stanja i na odluku modela učenja potkrepljivanjem da izabere određenu akciju, uzimajući u obzir sve moguće načine na koje se ta karakteristika može pojaviti u kombinaciji sa ostalim karakteristikama.

2.2.2 Veliki jezički modeli

Modelovanje jezika (eng. *Language Modeling*) imalo je ključnu ulogu u razvoju jezičke inteligencije mašina. Istraživanje u ovoj oblasti prošlo je kroz četiri razvojne faze (Hadi et al. 2023):

- **Prva faza** obuhvatala je statističke jezičke modele (eng. *Statistical Language Models*) (Jelinek 1998). Tipičan primer predstavljali su *n-gram* modeli, koji su procenjivali verovatnoću sledeće reči u sekvenici na osnovu učestalosti pojavljivanja prethodnih n reči (Gao and Lin 2004). Na primer, *bigram* model koristi učestalost pojavljivanja svakog para uzastopnih reči u korpusu kako bi procenio verovatnoću da se određena reč pojavi nakon date prethodne reči (Collins 1996).
- **Druga faza** obuhvatala je primenu jezičkih modela zasnovanih na neuronskim mrežama (eng. *Neural Language Models, NLM*) (Bengio et al., 2000). Ovi modeli koristili su neuronske mreže za procenu distribucije verovatnoće sledeće reči na osnovu prethodnih reči u sekvenci. Za razliku od *n-gram* modela, NLM su omogućavali generalizaciju na neviđene kombinacije reči, zahvaljujući učenju semantičkih reprezentacija (*embedding*) reči. Primeri uključuju rekurentne neuronske mreže (eng. *Recurrent neural networks*) (Kombrink et al. 2011) i njihove varijante.
- **Treća faza** obuhvatala je uvođenje kontekstualnih reprezentacija reči (eng. *contextualized word embeddings*), koje su omogućile da se značenje reči određuje dinamički, u zavisnosti od njenog konteksta unutar rečenice. Za razliku od statičkih *embeddinga*, ovaj pristup koristio je unapred trenirane jezičke modele (eng. *Pre-Trained Language Models*), čiji su parametri učeni na velikim korpusima. Jedan od najznačajnijih primera ove faze bio je BERT (eng. *Bidirectional Encoder Representations from Transformers*) (Devlin et al. 2019).

- **Četvrta faza** obuhvatala je pojavu velikih unapred treniranih LLM-ova, koji su demonstrirali sposobnost rešavanja širokog spektra zadataka iz oblasti obrade prirodnog jezika (eng. *Natural Language Processing*) (Wei et al. 2022) bez potrebe za dodatnim treniranjem za svaki pojedinačni zadatak.

Danas se LLM-ovi koriste za rešavanje širokog spektra zadataka, uključujući pisanje eseja, sumarizaciju teksta i automatski prevod (J. Yang et al. 2024) (Hadi et al. 2023). LLM-ovi nalaze primenu u brojnim domenima, među kojima su obrazovanje (Susnjak and McIntosh 2024), medicina (Thirunavukarasu et al. 2023), finansije (Wu et al. 2023) i inženjstvo (Badini et al. 2023).

Jedan od savremenih LLM-ova je ChatGPT, model razvijen od strane kompanije *OpenAI* (OpenAI 2022). Ovaj model pripada klasi jezičkih modela poznatoj kao unapred trenirani generativni transformeri (eng. *Generative Pre-trained Transformer*, GPT) (Fui-Hoon Nah et al. 2023). GPT modeli primenjuju tehnike dubokog učenja i treniraju se na velikim skupovima podataka tokom dužeg vremenskog perioda (Casella et al. 2023).

ChatGPT je unapred treniran veliki jezički model koji je dodatno prilagođen (eng. *fine-tuned*) za konverzijske zadatke, čime je optimizovan za generisanje koherentnih i prirodnih odgovora u interakciji sa korisnikom (Fui-Hoon Nah et al. 2023). Ovaj model pripada široj kategoriji generativne veštačke inteligencije (eng. *Generative AI*), koja obuhvata alate sposobne za automatsko generisanje teksta, kreiranje kreativnog sadržaja, kao i integraciju i interpretaciju informacija iz različitih izvora (Hadi et al. 2023) (Khoury et al. 2023) (Haase and Hanel 2023). U okviru ove disertacije koristi se ChatGPT model zasnovan na GPT-4 arhitekturi (Koubaa 2023).

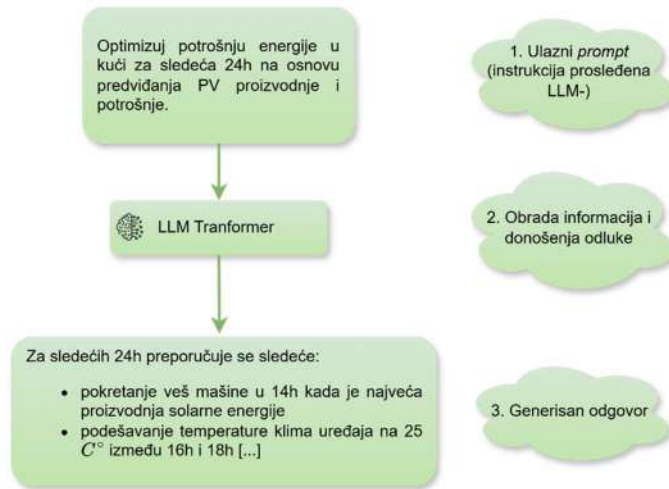
Slika 9 ilustruje proces kojim LLM generiše izlazni tekst (eng. *completion*) na osnovu ulaznog teksta (eng. *prompt*). Prvi korak obrade ulaznog teksta je njegova numerička reprezentacija. Tekst se najpre segmentira u tokene – u zavisnosti od primenjene metode segmentacije, tokeni mogu predstavljati reči, podreči, karaktere ili bitove. Svaki token se zatim mapira na jedinstveni numerički identifikator, koji odgovara njegovoj poziciji u rečniku tokena – skupu svih mogućih tokena koji model može prepoznati i obraditi. Ovi identifikatori se potom konvertuju u embedding vektor (eng. *embedding vector*), koji modelu omogućavaju da obrađuje značenje i odnose između tokena u višedimenzionom prostoru. Proces tokenizacije ulaznog teksta detaljnije je opisan u potpoglavlju 2.2.2.1.

Embedding vektori se zatim prosleđuju neuronskoj mreži zasnovanoj na transformer arhitekturi (Vaswani et al. 2023), koja je detaljnije prikazana u potpoglavlju 2.2.2.2. Transformeri se mogu koristiti za različite zadatke, u zavisnosti od konteksta primene. U ovom potpoglavlju opisan je generativni scenario, u kojem neuronska mreža na izlazu vraća distribuciju verovatnoća nad celokupnim rečnikom tokena, odnosno verovatnoće svih mogućih tokena u rečniku modela koji mogu slediti nakon ulaznog teksta. Proces izbora sledećeg tokena iz ove distribucije zavisi od primenjene strategije dekodiranja, koja je detaljnije opisana u potpoglavlju 2.2.2.3.

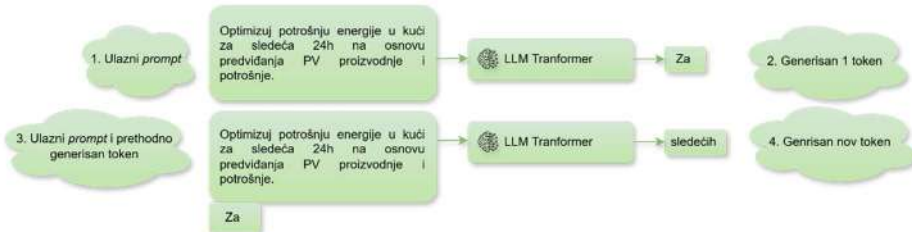
Slika 10 prikazuje proces generisanja izlaznog teksta, koji se odvija token po token. Svaki generisani token zahteva jedan prolazak unapred (eng. *forward pass*) ulaznog teksta kroz model. LLM-ovi su autoregresivni (eng. *autoregressive*), što znači da model, prilikom predikcije sledećeg tokena, na ulaz dodaje sve prethodno generisane tokene (Naveed et al. 2024). Ovaj proces se završava kada model generiše specijalni token za kraj sekvence

(eng. *end-of-sequence token*, EOS) ili kada dostigne maksimalan broj dozvoljenih tokena u izlaznom tekstu.

Da bi mogao da generiše koherentan i smislen tekst, model se prethodno obučava na velikim korpusima podataka. Jezički model koji je prošao kroz ovu fazu obuke naziva se pre-treniran veliki jezički model (eng. *pre-trained LLM*). Tokom obuke, model uči statističke obrasce jezika i formira reprezentacije značenja reči i fraza, što mu omogućava da kasnije predviđa sledeće tokene u različitim kontekstima. Dalji tekst će predstaviti ključne komponente transformer arhitekture koje omogućavaju ovaj proces.



Slika 9. Na visokom nivou apstrakcije, veliki jezički model prima ulazni tekst u obliku prompta (eng. *prompt*) i, nakon obrade, generiše izlazni tekst (eng. *completion*).



Slika 10. Izlazni tekst se generiše iterativno, token po token. Prvi izlazni token se generiše na osnovu zadatog prompta. Svaki naredni token generiše se tako što model koristi prompt proširen prethodno generisanim tokenima kao novi ulaz.

2.2.2.1 Tokenizator

Tokenizator je ključna komponenta LLM arhitekture, odgovorna za segmentaciju ulaznog teksta u niz tokena koje model može da obradi. U zavisnosti od izabrane strategije za tokenizaciju, tokeni mogu biti cele reči, podreči, ili pojedinačni karaktere. Ova odluka utiče na efikasnost modela: manji broj dužih tokena ubrzava obradu, ali umanjuje fleksibilnost za retke i nepoznate reči, dok finija segmentacija povećava broj tokena i troškove obrade, ali poboljšava robusnost.

Prilikom dizajniranja tokenizatora, definišu se i veličina rečnika i specijalni tokeni. Veličina rečnika određuje broj jedinstvenih tokena koje model može da prepozna i

generiše – veći rečnik omogućava veću pokrivenost jezika, ali povećava memorijske zahteve i složenost treniranja (Mikolov et al. 2013). Specijalni tokeni su unapred definisani simboli sa posebnim funkcijama unutar modela – na primer, označavanje početka ili kraja sekvence ili razdvajanje rečenica.

Tokenizator se zatim trenira (ili prilagođava) na reprezentativnom tekstualnom korpusu. Trening tokenizatora podrazumeva statističku analizu jezika – identifikaciju frekventnih obrazaca, konstrukciju rečnika i određivanje optimalne segmentacije. Ovaj proces omogućava da model kasnije efikasno transformiše tekst u niz tokena, prilagođen specifičnostima jezika, domena i ciljeva zadatka.

U ovoj disertaciji koristi se GPT-4, koji je ostvario značajno bolje rezultate u poređenju sa prethodnim generacijama GPT modela (Koubaa 2023) (OpenAI et al. 2024). U ovom istraživanju je ustanovljeno da su promptovi prikazani u potpoglavlju 4.5 previše složeni za starije verzije GPT modela, dok ih je GPT-4 uspešno obradio. GPT modeli koriste *BytePair Encoding* tokenizaciju (Sennrich et al. 2016). Ovaj tokenizator segmentira tekst na jedinice koje mogu predstavljati reči, podreči, pojedinačne znakove ili frekventne kombinacije karaktera i simbola. U slučaju GPT-4 modela, koristi se rečnik koji sadrži nešto više od 100,000 tokena. Primer rezultata tokenizacije GPT-4 modela na *OpenAI* platformi je prikazan na slici 11, gde je ilustrovano kako se ulazni tekst transformiše u niz tokena pre nego što se prosledi modelu na obradu.



Slika 11. Tokenizator, na osnovu specifične metode i prethodno definisane procedure treniranja, segmentira tekst na tokene pre nego što se tekst preda LLM modelu. (OpenAI 2022)

Nakon što tokenizator segmentira ulazni tekst i svakom tokenu dodeli jedinstveni numerički identifikator, ti identifikatori se mapiraju na embedding, kao što je prikazano na slici 12. Embedding je gusti (eng. *dense*) vektor realnih brojeva fiksne dimenzije, koja se unapred definiše kao hiperparametar modela. Tipične dimenzionalnosti embeddinga kreću se od 8 u jednostavnijim modelima, do 1024 kod velikih jezičkih modela. Veća dimenzionalnost omogućava modelu da preciznije modeluje semantičke odnose među tokenima, ali zahteva veću količinu podataka i računarske resurse za treniranje.

Embedding enkodira semantičke informacije o tokenu – slične reči su predstavljene vektorima sličnih vrednosti. U kontekstualnim modelima, kao što je GPT, isti token može imati različite embeddinge u zavisnosti od konteksta u kojem se javlja, čime se postiže fleksibilnije i tačnije razumevanje značenja. Nasuprot tome, modeli koji ne koriste kontekst (na primer, *word2vec*) dodeljuju svakoj reči jednu, fiksnu reprezentaciju, nezavisno od njenog konteksta u rečenici.

Embedding vektori se ne zadaju unapred, već se njihovi parametri uče tokom procesa treniranja modela – predstavljaju optimizovane težine neuronske mreže. Primer obučavanja embeddinga reči ilustrovan je kroz algoritam *word2vec* (Mikolov et al. 2013). Za njegovo treniranje koristi se tekstualni korpus koji se sastoji od skupa rečenica iz kojih se generišu parovi ulaz-izlaz za obuku modela. *Word2vec* primenjuje dva osnovna pristupa (Mikolov et al. 2013):

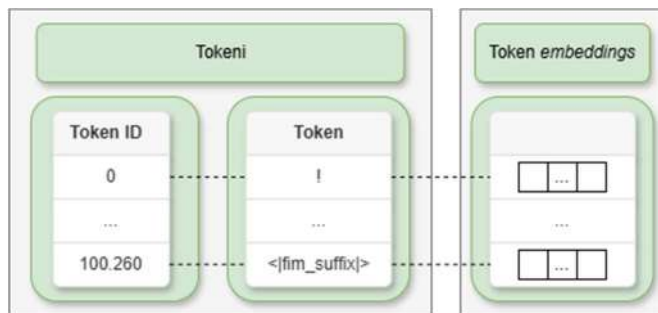
- *Continuous Bag of Words* (CBOW) – koristi reči iz konteksta kako bi prevideo ciljnu reč.
- *Skip-gram* – koristi ciljnu reč kako bi predvideo reči koje se pojavljuju u njenom kontekstu, tj. u neoprednoj blizini.

Primer 2.3 ilustruje razliku između CBOW i *Skip-gram* modela u načinu predviđanja reči na osnovu konteksta.

Primer 2.3 Postavka CBOW i *skip-gram* modela na primeru rečenice: “*Mačka sedi na prozoru i posmatra ptice.*” Ukoliko je ciljna reč “*prozoru*”, a veličina konteksta (eng. *window size*) 2, dobijamo sledeće:

- CBOW koristi reči iz konteksta: „*sedi*“, „*na*“, „*i*“, „*posmatra*“ kako bi prevideo ciljnu reč „*prozoru*“.
- *Skip-gram* koristi ciljnu reč „*prozoru*“ kao ulaz, i predviđa reči iz njenog konteksta: „*sedi*“, „*na*“, „*i*“, „*posmatra*“.

CBOW se pokazao efikasnijim za učestale reči i pogodniji je za treniranje na velikom skupovima podataka, dok *skip-gram* bolje modeluje retke reči i uspešnije uči semantičke relacije u manjim korpusima.



Slika 12. Sa leve strane prikazan je trenirani tokenizator, dok je sa desne strane prikazan LLM model koji sadrži embeddinge. Svaki token ima svoju odgovarajuću vektorsku reprezentaciju.

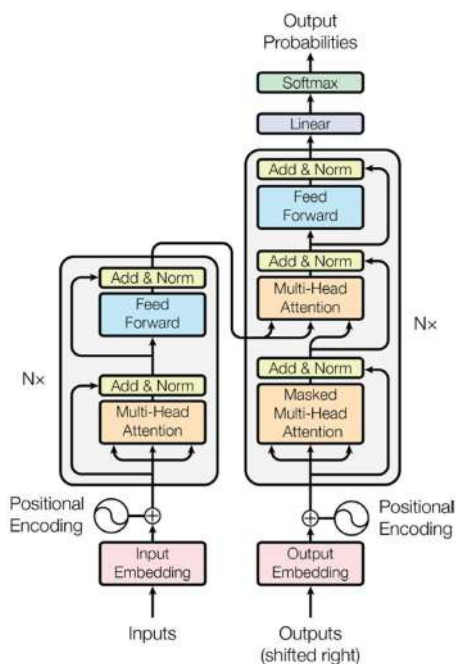
Kao rezultat opisanog procesa, na ulaz LLM modela dolazi niz embeddinga, koji predstavlja numeričku reprezentaciju teksta u višedimenzionalnom prostoru. Rezultujući

niz embeddinga se prosleđuju transformer arhitekturi, gde se obrađuju kroz slojeve neuronske mreže.

2.2.2.2 Transformer arhitektura

Detaljniji prikaz transformer arhitekture dat je na slici 13. Transformer se sastoji od sledećih komponenti:

- Pozicioni enkoding (eng. *positional encoding*) – dodaje informacije o redosledu tokena unutar ulaznih sekvenci, budući da transformer nema ugrađenu sekvencijalnu obradu (potpoglavlje 2.2.2.2.1).
- Enkoder (eng. *encoder*) – transformiše ulaznu sekvencu u kontekstualizovane reprezentacije (eng., *contextualized embedding*), omogućavajući modelu da razume odnose među tokenima. Za svaki ulazni token, enkoder generiše vektorsku reprezentaciju, koja se obogaćuje informacijama iz celokupnog konteksta sekvence (Vaswani et al. 2023).
- Dekoder (eng. *decoder*) – koristi izlaz enkodera i prethodno generisane tokene kako bi autoregresivno generisao sledeće tokene u izlaznoj sekvenci. Detaljna analiza međusobne interakcije enkodera i dekodera u procesu generisanja teksta data je u potpoglavlju 2.2.2.2.3.
- Mehanizam višestruke pažnje (eng. *multi-head attention*) – omogućava modelu da analizira odnose između tokena u sekvenci (potpoglavlje 2.2.2.2.2).



Slika 13. Transformer arhitektura (Vaswani et al. 2023).

2.2.2.2.1 Pozicioni enkoding

Ključna prednost transformera u odnosu na rekurentne modele je njihova sposobnost paralelnog procesiranja tokena unutar ulazne sekvence, što značajno povećava efikasnost obrade i treniranja modela (Lin et al. 2022). Efikasna paralelizacija omogućava treniranje

transformera nad izrazito velikim skupovima podataka, što ih čini pogodnim za složene zadatke obrade prirodnog jezika.

Međutim, zbog paralelne obrade, tranformeri nemaju inherentnu sposobnost razumevanja redosleda tokena u sekvenci, za razliku od rekurentnih modela koji prirodno obrađuju podatke sekvencijalno. Zbog toga je neophodno eksplicitno dodati informacije o poziciji svakog tokena u sekvenci, što se postiže pozicionim enkodiranjem (eng. *positional embedding*) (Vaswani et al. 2023).

Svaki token u ulaznoj sekvenci dobija svoj pozicioni enkoding, čija je dimenzija jednaka dimenziji njegovog embeddinga tokena (eng. *token embedding*). Pozicioni enkoding se zatim sabira sa embeddingom tokena, a dobijeni rezultat se prosleđuje kao ulaz u sve slojeve enkodera. Slično, dekodier koristi sumu pozicionih enkodinga i embeddinga svojih generisanih izlaznih tokena.

Postoji više načina za definisanje pozicionog enkodinga (Gehring et al. 2017). Autori rada (Vaswani et al. 2023) predložili su sinusno-kosinusni pozicioni enkoding, koji omogućava modelu da prepoznaje relativne odnose između tokena bez potrebe za dodatnom optimizacijom parametara. Ovaj metod koristi periodične funkcije za kodiranje pozicije tokena, čime se omogućava generalizacija na sekvence različitih dužina. Pozicioni enkoding se računa prema sledećim formulama:

$$PE_{(pos,2i)} = \sin\left(\frac{pos}{10000^{\frac{2i}{d_{model}}}}\right),$$

$$PE_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{\frac{2i}{d_{model}}}}\right),$$
(24)

gde je pos pozicija tokena u sekvenci, d_{model} dimenzija vektorskog prostora reprezentacija tokena, a i indeks dimenzije unutar vektora pozicionog enkodinga.

Primer 2.4 Izračunavanje pozicionog enkodinga analizirajući rečenicu: „*Mačka juri miša dok pas spava*“. Radi jednostavnosti, dimenzija vektorskog prostora postavljena je na $d_{model} = 4$, iako je u praksi ovaj broj znatno veći kako bi enkoding obuhvatio složenije semantičke odnose među tokenima. Indeks i je u opsegu $0 \leq i \leq \frac{d}{2}$. Tabela 4 prikazuje vrednosti pozicionih enkodinga svih reči u rečenici za ovaj slučaj, izračunate na osnovu jednačina (24). Posmatrajući vrednosti u tabeli 4, može se primetiti da svaki token dobija jedinstven pozicioni enkoding koji označava njegovu apsolutnu poziciju u sekvenci. Pozicioni enkoding omogućava modelu da prepozna relativne i apsolutne položaje reči, čime postaje sposoban za obavljanje zadataka kao što su prevođenje i generisanje teksta.

Tabela 4. Primer izračunatih vrednosti pozicionog enkodinga (zaokruženo na dve decimale).

Reč	Pozicija u rečenici (pos)	$PE_{(pos,0)}$	$PE_{(pos,1)}$	$PE_{(pos,2)}$	$PE_{(pos,3)}$
Mačka	0	0.00	1.00	0.00	1.00
juri	1	0.84	0.54	0.01	1.00
miša	2	0.90	-0.41	0.02	1.00
dok	3	0.14	-0.99	0.03	1.00
pas	4	-0.75	-0.65	0.04	0.99
spava	5	-0.95	0.28	0.05	0.99

2.2.2.2.2 Mehanizam pažnje

Mehanizam pažnje (eng. *attention*) omogućava modelima dubokog učenja da se fokusiraju na ključne delove ulazne sekvence prilikom obrade podataka. Tradicionalni sekvencijalni modeli, poput rekurentnih neuronskih mreža (eng. *Recurrent Neural Network*), obrađuju podatke iterativno, gde svaka nova reprezentacija zavisi od prethodnih stanja. Nasuprot tome, mehanizam pažnje omogućava modelu da istovremeno analizira sve delove sekvence, čime se poboljšava efikasnost i omogućava bolja dugoročna zavisnost među tokenima.

Mehanizam pažnje zasniva se na tri ključne komponente:

- Upit (eng. *query*, q) – vektor koji predstavlja trenutno stanje modela pri obradi određenog tokena u sekvenci. Svi upiti u jednoj sekvenci formiraju matricu upita (eng. *query matrix*, Q), gde svaki red predstavlja jedan vektor upita za pojedinačan token.
- Ključ (eng. *key*, k) – vektor koji kodira informacije o karakteristikama pojedinačnog tokena u sekvenci. Skup svih ključeva formira matricu ključeva (eng. *key matrix*, K), gde svaki red odgovara ključnom vektoru jednog tokena.
- Vrednost (eng. *value*, v) – vektor koji sadrži sadržajne informacije vezane za dati token. Svi vektori vrednosti čine matricu vrednosti (eng. *value matrix*, V), gde svaki red nosi reprezentaciju pojedinačnog tokena u sekvenci.

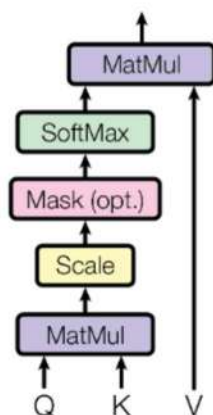
Mehanizam pažnje može se opisati kao funkcija koja, na osnovu matrice upita (Q) i skupa ključ-vrednost parova (K - V), računa izlazne vektore pažnje. Izlaz pažnje predstavlja ponderisanu sumu matrice vrednosti (V), pri čemu se težine pažnje određuju na osnovu sličnosti između matrice upita (Q) i matrice ključeva (K).

Sličnost između upita i ključeva računa se kao proizvod QK^T , pri čemu svaka vrednost u rezultujućoj matrici predstavlja meru sličnosti između jednog upitnog vektora i jednog ključnog vektora. Da bi se poboljšala stabilnost tokom obuke modela, dobijene vrednosti se skaliraju deljenjem sa $\sqrt{d_k}$, gde d_k predstavlja dimenziju vektora ključeva. Konačno, rezultati se normalizuju *softmax* funkcijom, koja transformiše sličnosti u distribuciju verovatnoća. Ovaj pristup je poznat kao pažnja zasnovana na skaliranom skalarnom proizvodu (eng. *Scaled Dot-Product Attention*) (slika 14).

Matematički, pažnja se računa po sledećoj formuli (Vaswani et al. 2023):

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V. \quad (25)$$

Ova formula omogućava modelu da dodeli veću pažnju ključevima koji su relevantniji za dati upit, čime se model fokusira na značajne informacije u sekvenci i poboljšava razumevanje konteksta.



Slika 14. Pažnja zasnovana na skaliranom skalarnom proizvodu vektora.

U standardnoj transformer arhitekturi postoje dva glavna tipa pažnje:

- Samo-pažnja (eng. *self-attention*) – koristi se kada model analizira odnose unutar iste sekvence. Enkoder koristi samo-pažnju kako bi izgradio reprezentaciju na osnovu međusobnih odnosa reči u ulaznoj sekvenci. Dekoder koristi samo-pažnju kako bi procenio odnose među već generisanim tokenima. U primeru 2.5 prikazana je upotreba mehanizma samo-pažnje.
- Enkoder-dekoder pažnja (eng. *encoder-decoder attention*) – koristi se u dekoderu za interakciju sa izlaznim reprezentacijama enkodera. Omogućava dekoderu da generiše sledeći token oslanjajući se na informacije iz enkodera.

Primer 2.5 U samo-pažnji, model analizira koliko je svaki token u sekvenci povezan sa ostalim tokenima. Posmatrajmo jednostavnu rečenicu: „Mačka juri miša dok pas spava“. Prvo, svaka reč se transformiše u vektorski embedding, a zatim se za svaku od njih izračunavaju odgovarajući upiti (Q), ključevi (K) i vrednosti (V):

- Matrica upita (Q) – svaki njen red je vektor upita koji model koristi da bi odredio koliko su ostali tokeni važni za taj konkretan token.
- Matrica ključeva (K) – svaki red u ovoj matrici predstavlja vektor koji sadrži informacije o specifičnom tokenu, što omogućava modelu da poredi token sa ostalim upitima.
- Matrica vrednosti (V) – svaki red nosi sadržajnu reprezentaciju tokena, koja se kasnije ponderiše i koristi za računanje izlaza pažnje.

Na primer, ako se fokusiramo na reč „juri“, model računa kompatibilnost njenog upitnog vektora sa svim ključnim vektorima u sekvenci pomoću proizvoda QK^T . Ovo daje matricu pažnje, gde svaka vrednost pokazuje stepen sličnosti između reči „juri“ i svih ostalih reči u rečenici.

Pretpostavimo da je model dodelio sledeće težine pažnje za „juri“ u kontekstu ove rečenice: „mačka“ $\rightarrow 0.1$, „juri“ $\rightarrow 0.2$, „miša“ $\rightarrow 0.7$, „dok“, „pas“, „spava“ $\rightarrow \approx 0.0$. Najveću pažnju (0.7) „juri“ pridodaje reči „miša“ jer su semantički povezane – glagol

„juri“ se odnosi na objekat „miša“. Takođe, „juri“ obraća određenu pažnju na „mačka“, jer ona vrši radnju jurenja. Preostale reči imaju gotovo zanemarljive težine, jer nisu relevantne za akciju „juri“ u datom trenutku obrade.

Ulazna sekvenca prvo prolazi kroz ulazni embedding sloj i sloj pozicionog enkodiranja, čime se za svaki token generiše vektorska reprezentacija koja sadrži informacije o značenju reči i njenom položaju u rečenici. Dobijene vektorske reprezentacije zatim ulaze u mehanizam samo-pažnje, gde se transformišu u tri matrice: upite (Q), ključeve (K) i vrednosti (V). Ove transformacije se ostvaruju primenom linearnih projekcija, pri čemu se svaka ulazna vektorska reprezentacija X množi sa odgovarajućom težinskom matricom:

$$Q = XW^Q, K = XW^K, V = XW^V.$$

Ovde je:

- X matrica dimenzija $n \times d_{model}$, gde je n broj tokena, a d_{model} dimenzija vektorske reprezentacije,
- W^Q , W^K i W^V su parametarske matrice koje se optimizuju tokom obuke modela.

Transformisane reprezentacije Q, K i V omogućavaju modelu da izračuna težine pažnje, odnosno da proceni koliko je svaki token relevantan u odnosu na ostale u sekvenci. Ove težine se koriste za izgradnju kontekstualizovanih reprezentacija, koje uzimaju u obzir međusobne zavisnosti između tokena.

U višeslojnoj arhitekturi transformera, svaki sledeći enkoderski sloj dodatno obogaćuje reprezentacije reči na osnovu pažnje iz prethodnog sloja. Na taj način, model progresivno uči da prepoznaje dublje semantičke i sintaksičke odnose između reči u rečenici. Mehanizam pažnje omogućava modelu da adaptivno raspodeli pažnju na različite delove sekvence, u skladu sa njihovim značenjem u datom kontekstu.

Višestruki mehanizam pažnje (eng. *multi-head attention*), prikazan na slici 15, predstavlja proširenje osnovnog mehanizma pažnje kroz istovremenu primenu više nezavisnih pažnji. Ova tehnika omogućava modelu da se istovremeno fokusira na različite odnose unutar sekvence, čime se značajno povećava njegova sposobnost razumevanja složenih zavisnosti u tekstu. Svaka pažnja u okviru višestrukog mehanizma nezavisno analizira različite karakteristike podataka, nakon čega se njihovi izlazi kombinuju u jednu bogatiju reprezentaciju.

Matematički, višestruka pažnja se opisuje sledećom formulom (Vaswani et al. 2023):

$$MultiHead(Q, K, V) = Concat(head_1, \dots, head_h)W^O,$$

gde je:

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V). \quad (26)$$

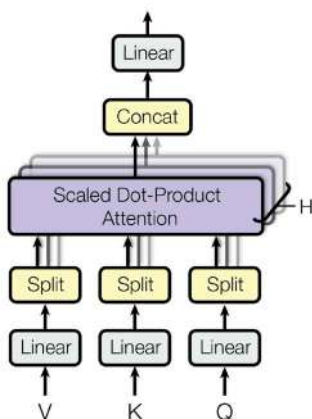
U ovoj formulaciji:

- Q, K i V su ulazne matrice koje sadrže upite, ključeve i vrednosti za celu sekvencu.
- W_i^Q, W_i^K i W_i^V su matrice projekcije, koje transformišu ulazne vektore Q, K i V u različite *podprostore*, omogućavajući svakom mehanizmu pažnje da analizira

različite aspekte podataka. Drugim rečima, svaka glava pažnje koristi sopstvenu verziju ulaznih podataka, dobijenu linearnim transformacijama pomoću ovih matrica.

- Svaka glava pažnje ($head_i$) primenjuje standardni mehanizam pažnje (25), ali nad različitim transformisanim verzijama ulaznih matrica Q , K i V , dobijenim pomoću matrica projekcije W_i^Q , W_i^K i W_i^V .
- *Concat* označava spajanje svih pojedinačnih izlaza pažnji u jedan objedinjeni vektor.
- W^O je matrica težina, koja transformiše spojen izlaz višestruke pažnje u odgovarajući prostor dimenzija, pripremajući ga za sledeće slojeve mreže.

Korišćenjem višestruke pažnje, model postiže veću izražajnu moć, jer svaka glava može da se fokusira na različite odnose između reči u rečenici. Na primer, dok jedna glava pažnje može da identifikuje subjekat i glagol, druga može da poveže prideve sa imenicama, a treća može da prepozna dugoročne zavisnosti u rečenici.



Slika 15. Višestruki mehanizam pažnje (Vaswani et al. 2023).

2.2.2.2.3 Enkoder i dekode

Primarna funkcija enkoder komponente jeste transformacija ulazne sekvence u kontekstualizovane vektorske reprezentacije korišćenjem višestrukog mehanizma pažnje. U originalnom radu (Vaswani et al. 2023), enkoder se sastojao od šest identičnih slojeva. Svaki sloj uključivao je dva glavna podsloja: višestruki mehanizam pažnje i potpuno povezanu neuronsku mrežu (eng. *feedforward neural network*). Nakon svakog podsloja primenjivane su rezidualne konekcije i sloj za normalizaciju (slika 13).

Glavna uloga dekodera jeste autoregresivno generisanje izlazne sekvence, pri čemu koristi prethodno generisane tokene i kontekstualizovane reprezentacije iz enkodera kako bi predvideo sledeći token. U radu (Vaswani et al. 2023), dekode se takođe sastojao od šest identičnih slojeva. Svaki sloj obuhvatao je tri podsloja: prva dva identična onima u enkoderu, dok je treći bio višestruki mehanizam pažnje nad izlazima enkodera, čime se omogućava uvođenje informacija iz ulazne sekvence u proces generisanja (slika 13). Kao i kod enkodera, svi podslojevi bili su povezani rezidualnim vezama i normalizovani posebnim slojevima.

Za razliku od enkodera, dekodер koristi maskirani višestruki mehanizam pažnje (eng. *masked multi-head attention*) kako bi sprečio da budući tokeni utiču na generisanje trenutnog tokena. Time se obezbeđuje da se prilikom treniranja svaki token predviđa isključivo na osnovu prethodno poznatih tokena.

2.2.2.2.4 Strategije izbora sledećeg tokena

Odabir sledećeg tokena zasniva se na strategiji dekodiranja (eng. *decoding strategy*). Strategije dekodiranja mogu se podeliti na (Shi et al. 2024):

- Determinističke strategije – isti ulaz uvek rezultira istim izlazom.
- Stohastičke strategije – uvode nasumičnost, što može dovesti do varijabilnosti izlaza čak i za isti ulaz.

Najjednostavnija deterministička strategija je pohlepna strategija (eng. *greedy strategy*), pri kojoj se u svakom koraku bira token sa najvećom verovatnoćom. Iako je ovaj pristup računski efikasan, često dovodi do repetitivnih ili suboptimalnih izlaza. Stohastičke strategije uvode nasumičnost kako bi poboljšale varijabilnost i kreativnost generisanog teksta. Ove metode omogućavaju izbor tokena iz skupa najverovatnijih kandidata, umesto strogo najverovatnijeg tokena, čime se smanjuje rizik od predvidljivih ili monotonih izlaza.

Jedan od najčešće korišćenih načina za kontrolisanje nasumičnosti u dekodiranju je uzorkovanje sa temperaturom (eng. *temperature sampling*). Ovaj postupak podrazumeva postavku parametra temperature (eng. *temperature*), čija vrednost može biti bilo koji realan broj ≥ 0 , iako se u praksi najčešće definiše u opsegu $[0, 1]$ (Shi et al. 2024). Vrednosti temperature bliže 1 ukazuju na smanjenje preciznosti LLM modela (Renze 2024). Niže vrednosti temperature (bliže 0) vode ka determinističkom odabiru tokena, slično pohlepnoj strategiji. Više vrednosti temperature povećavaju nasumičnost u izboru tokena, što dovodi do raznovrsnijih izlaza, ali može smanjiti i koherentnost generisanog teksta. Parametar temperature određuje balans između predvidljivosti i varijabilnosti, što je ključno za prilagođavanje izlaznog teksta različitim aplikacijama.

2.2.2.3 Tehnika prompt inženjeringa

Veliki jezički modeli koriste se za rešavanje zadataka opšte namene (Zhao et al. 2024). Njihova primena je posebno izražena u interakciji sa korisnicima putem prirodnog jezika, gde generišu odgovore koji su stilom i strukturom slični ljudskoj komunikaciji. Među poznatim primerima ovih modela su GPT-3, GPT-4, ChatGPT i Claude (OpenAI 2022) (Claude, 2023).

Za komunikaciju sa modelom koristi se prompt (Liu et al. 2023), koji predstavljaju instrukciju, pitanje ili opis problema namenjen usmeravanju modela ka generisanju tačnih i relevantnih odgovora (White et al. 2023). Kvalitet prompta značajno utiče na performanse modela, jer određuje način na koji model tumači zadatak i oblikovati izlaz (He et al. 2024).

Tehnika prompt inženjeringa (eng. *prompt engineering*) predstavlja jedan od načina prilagođavanja LLM-a za rešavanje specifičnih zadataka (Liu et al. 2023). Ovaj proces podrazumeva osmišljavanje i optimizaciju prompta, napisanog prirodnim jezikom, sa ciljem usmeravanja modela ka generisanju odgovora prilagođenih specifičnim zahtevima.

Struktura prompta može biti jednostavna – nekolicina reči ili rečenica (slika 16). Dok su jednostavni promptovi dovoljni za dobijanje osnovnih informacija, složeniji zadaci često zahtevaju preciznije definisane instrukcije koje modelu pružaju dodatni kontekst i jasna očekivanja u vezi sa formatom odgovora.



Slika 16. Jednostavni prompt bez instrukcija gde LLM treba da dovrši započetu rečenicu.

Jedan od najčešćih oblika prompt inženjeringa jeste prompt zasnovan na instrukcijama (eng. *instruction-based prompting*) (Efrat and Levy 2020) (Mishra et al. 2022). Ovaj pristup podrazumeva precizno definisanje konkretnog problema koji treba rešiti, uz jasno postavljene ciljeve. Ovaj pristup se koristi u različitim zadacima, uključujući nadglednu klasifikaciju (eng. *supervised classification*), pretragu informacija, sumarizaciju teksta, generisanja koda i prepoznavanje imenovanih entiteta (eng. *named entity recognition*). Svaki od ovih zadataka zahteva specifičan format *prompt*-a kako bi se postigao optimalan rezultat.

Formulisanje efektivnog prompta predstavlja iterativan proces koji zahteva primenu ključnih tehnika, poput preciznog definisanja zadatka, postavljanja univerzalnog odgovora za slučajeve kada LLM generiše halucinacije (Ji et al. 2023) i pažljivog strukturiranja redosleda instrukcija, pri čemu se preporučuje da se ključne instrukcije smeštaju na početak ili kraj prompta.

Postoje napredne tehnike prompt inženjeringa koje olakšavaju pisanje kompleksnijih *prompt*-ova. Dok se jednostavniji promptovi uglavnom sastoje od tri komponente – instrukcije, podataka i formata odgovora – kompleksniji promptovi mogu uključivati dodatne elemente, koji omogućavaju usmeravanje modela i poboljšavaju relevantnost generisanih odgovora. Kompleksniji promptovi, osim tri komponente koje sadrže jednostavniji promptovi, mogu uključiti i sledeće elemente:

- *Persona*: Definisanje specifične uloge LLM-a. Na primer, ukoliko se jasno precizira da LLM treba da simulira sistem za upravljanje pametnim uređajima sa ciljem optimizacije potrošnje energije i očuvanje udobnosti stanara, model će se u većini slučajeva usmeriti na relevantne aspekte zadatka (Shanahan et al.

2023) (Buren 2023). Postavljanje uloge pomaže modelu da bolje razume kontekst i prioritete, što utiče na preciznost i korisnost odgovora.

- **Publika:** Identifikacija ciljne grupe korisnika koji će tumačiti odgovor LLM-a. Na primer, u okviru problema upravljanja energijom u pametnoj kući, može se naglasiti da preporuke treba prilagoditi stanaru bez tehničkog predznajanja.
- **Ton:** Zadavanje tona odgovora, kao što su formalni ili neformalni, u zavisnosti od specifične situacije i publike. Na primer, ako se stanaru daju saveti o uštedi energije, može se zahtevati jasan i prijateljski ton, bez stručnih izraza, kako bi preporuke bile lako razumljive.
- **Podaci:** Relevantne informacije specifične za zadatak, koje mogu unaprediti kvalitet odgovora. Na primer, stanar može dopuniti prompt podacima o upotrebi uređaja u protekloj nedelji, čime se omogućava preciznija personalizacija odgovora.

Kompleksni promptovi imaju modularnu strukturu, što omogućava fleksibilnost u njihovom oblikovanju. Komponente se mogu dodavati ili uklanjati proizvoljno, a redosled njihovog navođenja se može prilagođavati specifikaciji zadatka. Slika 17 prikazuje primer prethodno opisanih komponenti i ilustruje kako se modularna struktura prompta može koristiti za optimizaciju odgovora modela. Proces kreiranja prompta je eksperimentalan i iterativan – na osnovu odgovara koji generiše LLM, prompt se prilagođava sve dok se ne pronade optimalna struktura za dati problem.

The image shows a vertical stack of green boxes, each representing a component of a modular prompt. The components are as follows:

- Persona:** Oponašaj model za obradu prirodnog jezika.
- Instrukcija:** Klasifikuj komentare prema sentimentu kao pozitivne, negativne ili neutralne.
- Kontekst:** Komentari predstavljaju korisničke povratne informacije o proizvodu.
- Format:** Odgovor treba da bude organizovan u tabelarnoj formi, gde svaka kolona prikazuje komentar i pripadajuću klasu.
- Publika:** Klasifikator će koristiti analitičari koji proučavaju korisničke povratne informacije.
- Ton:** Formulacije odgovora treba da budu formalne i profesionalne.
- Podaci:** Komentari: [...]

Slika 17. Primer kompleksnog prompt koji sadrži više komponenti.

Pored instrukcija, prompt se može proširiti konkretnim primerima odgovora, čime se dodatno pojašnjava očekivani ishod prompta. Ovaj postupak, poznat kao učenje iz konteksta (eng. *in-context learning*) (Brown et al. 2020), omogućava LLM modelima da bolje razumeju i rešavaju zadatke na osnovu datih primera, bez potrebe za dodatnom obukom. U zavisnosti od broja primera u promptu, razlikuju se pristupi: *zero-shot* (bez primera), *one-shot* (sa jednim primerom) i *few-shot* (sa nekoliko primera). U zavisnosti od broja primera u promptu, razlikuju se tri pristupa:

- *Zero-shot* – model dobija samo instrukciju, bez primera odgovora i oslanja se isključivo na prethodno stečeno znanje. Na primer, u kontekstu HEMS-a, *zero-shot* prompt može sadržati instrukciju da model treba da optimizuje potrošnju energije na osnovu prognoze sunčanog dana, koristeći postojeću bazu znanja.
- *One-shot* – modelu se pruža jedan primer odgovora kao referenca.
- *Few-shot* – model dobija nekoliko primera odgovora kako bi bolje uočio obrasce i prilagodio svoj izlaz. Na primer, u kontekstu HEMS-a, *few-shot* prompt može uključivati primere potrošačkih obrazaca iz prethodnih sličnih situacija, praćene očekivanim odgovorom sistema. Odgovor sistema može biti optimalna akcija ili preporuka koju sistem treba da pruži, uzimajući u obzir specifičan ulazni scenario.

Ove strategije omogućavaju primenu velikih jezičkih modela u specifičnim kontekstima, bez potrebe za dodatnim finim podešavanjem modela.

2.3 Analiza odluka modela zasnovanih na učenju potkrepljivanjem

SHAP (*SHapley Additive exPlanations*) biblioteka može se koristiti za analizu uticaja svake ulazne karakteristike trenutnog stanja na akciju koju model učenja potkrepljivanjem procenjuje kao najpovoljniju u datom stanju. SHAP takođe omogućava uvid u relativni značaj pojedinačnih karakteristika, kao i u način na koji model koristi njihove međusobne interakcije prilikom donošenja odluka. Ovaj pristup se preporučuje isključivo za interpretaciju ponašanja modela, odnosno za analizu korelacionih odnosa unutar naučenog modela. Važno je napomenuti da se SHAP uvidi ne bi trebali tumačiti kao uzročni efekti u stvarnom svetu, jer se zasnivaju na strukturi modela i statističkim zavisnostima u podacima, a ne na eksperimentalno potvrđenim uzročno-posledničnim vezama (Molnar et al. 2021).

SHAP vrednosti zasnivaju se na teoriji kooperativnih igara (eng. *game theory*), u kojoj se doprinos svake ulazne karakteristike posmatra analogno doprinosu pojedinačnog igrača u postizanju zajedničkog cilja. U kontekstu mašinskog učenja, to znači da se procenjuje koliki doprinos svaka karakteristika ima u ukupnoj predikciji modela, uzimajući u obzir sve moguće kombinacije prisustva i odsustva drugih karakteristika.

U kontekstu učenja potkrepljivanjem, SHAP vrednosti se računaju za konkretnu akciju koju je model izabrao u datom stanju. Pozitivna SHAP vrednost ukazuje na to da data karakteristika povećava verovatnoću predikcije te izabrane akcije, dok negativna vrednost označava da karakteristika imaju smanjujući efekat. Intenzitet uticaja određuje se apsolutnom vrednošću SHAP koeficijentata, čime se kvantifikuje značaj pojedinačnih karakteristika u procesu donošenja odluke. Na primer, u *Gridworld* okruženju, gde model

učenja potkrepljivanjem treba da nauči kako da pronađe put do cilja, SHAP analiza može pokazati da horizontalna koordinata cilja ima visoku pozitivnu SHAP vrednost, jer snažno utiče na izbor akcije „idi desno“ – na primer, kada je cilj desno od trenutne pozicije agenta. Istovremeno, vertikalna koordinata cilja u tom istom stanju može imati neutralan ili negativan doprinos, ukoliko nije relevantna za odabir te konkretne akcije. Ovakva analiza omogućava bolje razumevanje koje informacije iz stanja najviše doprinose odlukama modela, nezavisno od kompleksnosti modela.

U radu (Jiang et al. 2022), PPO model je treniran da postigne energetska efikasnost prilikom vožnje kroz raskrsnice sa adaptivnim semaforiskim sistemom, optimizujući ubrzanje vozila u odnosu na dinamičko stanje okruženja. Autori su koristili SHAP za interpretaciju procesa donošenja odluka u svakom vremenskom koraku vožnje, omogućavajući kvantifikaciju doprinosa svake ulazne karakteristike stanja na izbor akcije. Rezultati su vizuelizovani grafovima koji prikazuju pozitivne i negativne uticaje pojedinačnih karakteristika, kao što su udaljenost od semafora, trenutna faza svetla i preostalo vreme do promene faze. Na taj način, ponašanje modela postaje transparentnije, a strategija verodostojnija.

Autori (Beechey et al. 2023) prvi su pružili teorijsku analizu primene SHAP metode za objašnjavanje odluka modela učenja potkrepljivanjem. U svojoj analizi ukazali su na ključne nedostatke postojećih pristupa koji primenjuju SHAP za analizu vrednosne funkcije ili politike, ističući da takvi pristupi ne omogućavaju pouzdanu interpretaciju doprinosa pojedinačnih karakteristika stanja na performanse agenta. Kao alternativu, autori predlažu metodu zasnovanu na *Shapley* vrednostima za objašnjavanje performansi modela učenja potkrepljivanjem (eng. *Shapley Values for Explaining Reinforcement Learning-Performance*), koja umesto analize pojedinačnih predikcija razmatra promene u ukupnoj nagradi kada se agentu uskraćuju pojedine karakteristike stanja. Iako ovaj pristup omogućava dublji uvid u uzročne odnose između karakteristika stanja i performansi agenta, njegova primena zahteva redefinisavanje politike i uvodi dodatnu metodološku složenosti. Zbog toga nije korišćen u okviru ove disertacije, ali se prepoznaje kao važan pravac za buduća istraživanja u okviru interpretabilnosti modela učenja potkrepljivanjem.

3 Pregled aktuelnog stanja u oblasti

U ovom poglavlju analizira se postojeća literatura, sa fokusom na ključne aspekte razvoja i primene sistema za upravljanje energijom u pametnim kućama. Potpoglavlje 3.1 pruža istorijski pregled razvoja sistema za upravljanje energijom, prikazujući ključne tehnologije i pristupe kroz vreme. Potpoglavlje 3.2 analizira savremene strategije učenja potkrepljivanjem za upravljanje uređajima u pametnim kućama, sa posebnim naglaskom na primeni DDQN i PPO algoritama, koji su pokazali značajne rezultate. Potpoglavlje 3.3 razmatra primenu velikih jezičkih modela u optimizaciji potrošnje energije, pri čemu se ističe nedostatak literature o njihovoj primeni u automatizaciji upravljanja uređajima u pametnim kućama. Potpoglavlje 3.4 sumira ograničenja analiziranih rešenja, ističući otvorena istraživačka pitanja i obrazlažući potrebu za istraživanjem sprovedenim u ovoj tezi.

3.1 Istorijski pregled razvoja sistema za upravljanje energijom u pametnim kućama

Možemo tvrditi da termostati predstavljaju prvi oblik automatizovanog sistema za upravljanje energijom u domaćinstvu (Liu et al. 2016). Razvoj tehnologije i napredak računarskih sistema omogućili su pojavu softverskih rešenja za upravljanje električnom energijom.

Jedno od prvih takvih rešenja razvili su *Cook* i saradnici (Cook et al. 2003), koristeći inteligentnog agenta za automatizaciju upravljanja kućnim uređajima. Cilj ovog sistema bio je poboljšanja komfora stanara i povećanje energetske efikasnosti. Optimizacija se postizala algoritmima koji su predviđali sekvencu događaja u domaćinstvu, omogućavajući agentu da automatski reaguje na aktivnosti stanovnika i prilagođava se njihovim rutinama. Agent je koristio algoritme za analizu učestalosti događaja kako bi predvideo i automatski izvršavao zadatke, kao što su podešavanje temperature, uključivanje svetla i kafe aparata, kao i naručivanje namirnica. Na taj način ostvarena je ušteda energije i povećan komfor stanovnika. Važno je napomenuti da se u novijim istraživanjima određeni kućni uređaji, koji su u ovom sistemu bili automatizovani, sada klasifikuju kao „nekontrolisani potrošači“. To znači da njihov rad nije više pod kontrolom automatizovanog HEMS-a, već je ostavljen na upravljanje stanovnicima.

Sa porastom interesovanja za obnovljive izvore energije, istraživači su počeli da ih integrišu u HEMS rešenja kako bi poboljšali energetske efikasnosti. *Boynuegri* i saradnici (Boynuegri et al. 2013) razvili su HEMS zasnovan na jednostavnim „ako-onda“ (eng. *if-then*) pravilima. Sistem je uključivao obnovljive izvore energije, baterije za skladištenje energije i klima uređaj. Korišćenje klima uređaja bilo je ograničeno na letnji period i aktiviralo se isključivo tokom ciklusa hlađenja. Nisu razmatrani načini grejanja tokom hladnijih meseci, što je moglo umanjiti efikasnost sistema u zimskom periodu, kada su energetske potrebe i obrasci potrošnje drugačiji. Dodatno, pristup zasnovan na „ako-onda“ pravilima je neskalabilan jer zahteva definisanje specifičnih uslova za brojne moguće scenarije, uključujući varijacije u spoljnoj temperaturi, solarnoj proizvodnji, tarifnim režimima i karakteristikama uređaja. Rezultujući broj pravila može postati velik i nepregledan, što otežava održavanje, modifikaciju, proširenje i otklanjanje grešaka.

Jedno od najčešće korišćenih rešenja za energetske upravljanje zgradama jeste sistem za grejanje, ventilaciju i klimatizaciju (eng. *Heating, Ventilation, and Air-Conditioning*, HVAC). U literaturi je zabeležen značajan napredak u razvoju algoritama zasnovanih na učenju potkrepljivanjem za optimizaciju HVAC sistema. Yu i saradnici (Yu et al. 2020) implementirali su algoritam dubokih determinističkih gradijenata politike (eng. *Deep Deterministic Policy Gradients*, DDPG) za optimizaciju HVAC sistema, nekontrolisanih potrošača i sistema za skladištenje energije (eng. *Electrical Energy Storage*), poput baterija. Iako su postigli poboljšanja u termalnom komforu i energetskej efikasnosti, njihova studija bila je ograničena na letnje mesece, što može uticati na generalizabilnost rezultata na druge sezone.

Jedan od inovativnih koncepta za pametne kuće jeste integracija računarskog vida (eng. *Computer Vision*) s ciljem optimizacije potrošnje energije i poboljšanja komfora stanara. Jaihar i saradnici (Jaihar et al. 2020) implementirali su sistem koji je, pored ručnog upravljanja uređajima i automatizovane kontrole zasnovane na analizi korisničkih navika, koristio računarski vid za automatsko prepoznavanje emocionalnog stanja stanovnika. Na osnovu prepoznatih emocija, sistem je automatski prilagođavao osvetljenje i rad ventilatora. Ograničenje ove studije bilo je u tome što je eksperiment sproveden u kontrolisanom okruženju sa samo jednom osobom ispred kamere, što je ograničilo mogućnost generalizacije rezultata na realne uslove u kojima se u prostoriji može nalaziti više osoba. Takođe, studija nije pružila detaljan opis načina na koji su identifikovane emocije mapirane na rad uređaja. Korišćeni pristup bio je zasnovan na modelima mašinskog učenja za detekciju osam različitih emocija, ali eksperimentalna postavka, uključujući senzore i zvučne signalizatore, nije bila detaljno dokumentovana, što otežava razumevanje i praktičnu primenu dobijenih rezultata.

Nedavni pregled literature koji su sprovveli Gomes i saradnici (Gomes et al. 2022) utvrdio je da primena učenja potkrepljivanjem predstavlja najsavremeniju tehniku za automatizaciju upravljanja uređajima u pametnim kućama. Zbog toga je ovaj metodološki pristup usvojen kao osnova za ovo istraživanje.

3.2 Savremene strategije učenja potkrepljivanjem za upravljanje energijom u pametnim kućama

Ovo potpoglavlje pruža pregled razvoja tehnika učenja potkrepljivanjem za automatizovane sisteme upravljanja u pametnim kućama (HEMS).

Jedan od osnovnih pristupa u učenju potkrepljivanjem jeste Q-učenje (eng. *Q-Learning*), pri čemu njegova najjednostavniju varijantu predstavlja tabelarno Q-učenje (eng. *Tabular Q-Learning*). Xu i saradnici (Xu et al. 2021) primenili su tabelarno Q-učenje za binarnu kontrolu električnog bojlera s ciljem smanjenja potrošnje električne energije u domaćinstvu. Funkcija nagrade u ovom modelu zasnivala se na dva glavna faktora:

- Troškovima električne energije, koji su izračunavani na osnovu aktuelnih tarifa.
- Visokim negativnim penalima dodeljenim zbog odstupanja temperature vode od zadatih vrednosti u trenucima kada korisnici imaju potrebu za toplom vodom.

Ova metodologija omogućila je efikasno upravljanje energijom prilagođavanjem rada bojlera stvarnim potrebama stanovnika, čime su optimizovani potrošnja energije i komfor korisnika.

Xu i saradnici (Xu et al. 2020) takođe su koristili tabelarno Q-učenje. Za razliku od prethodnog rada, proširili su model dodavanjem novih tipova uređaja i uveli varijabilnu potrošnju energije, umesto pretpostavke da uređaji troše konstantnu količinu energije tokom rada. Modelovanje varijabilne potrošnje omogućilo je veću kontrolu rada uređaja. Pored toga, autori su definisali više-ciljnu funkciju nagrade, koja obuhvata energetske troškove i komfor korisnika, zavisno od tipa uređaja. Uređaji su klasifikovani u četiri kategorije:

- Uređaji sa fiksnim rasporedom rada (eng. *non-shiftable*),
- Uređaji sa prilagodljivom snagom rada (eng. *power-shiftable*),
- Uređaji čiji se rad prilagođava vremenskim tarifama (eng. *time-shiftable*),
- Električno vozilo, tretirano kao poseban tip uređaja.

Za sve tipove uređaja, osim onih sa fiksnim rasporedom rada, definisani su koeficijenti nezadovoljstva (eng. *dissatisfaction coefficient*) čije vrednosti nisu dodatno pojašnjene od strane autora. Koeficijenti nezadovoljstva se razlikuju u zavisnosti od tipa uređaja:

- Za uređaje sa promenljivom snagom rada, komfor se računao kao proizvod odgovarajućeg koeficijenta nezadovoljstva i kvadrata razlike između maksimalne potrebne energije i trenutne potrošnje uređaja.
- Za uređaje čiji se rad prilagođava vremenskim tarifama, komfor se definisao kao proizvod odgovarajućeg koeficijenta nezadovoljstva i kvadrata razlike između stvarnog i željenog vremena početka rada uređaja.
- Za električno vozilo, komfor je definisan kao proizvod odgovarajućeg koeficijenta nezadovoljstva i kvadratne razlike između maksimalne i trenutne energije za punjenje.

Za sve uređaje, osim električnog vozila, trošak električne energije računat je kao proizvod predviđene tarife i razlike između potrošnje energije uređaja i energije proizvedene iz solarnog panela. Kod električnog vozila, trošak se računao samo na osnovu potrošnje energije potrebne za punjenje. Na kraju, sve pojedinačne nagrade sabirane su kako bi se dobila ukupna funkcija nagrade sistema.

Glavni nedostatak tabelarnog Q-učenja jeste njegova zavisnost od Q-tabele fiksne veličine, koja služi za čuvanje i ažuriranje q-vrednosti za svaku moguću kombinaciju stanja i akcija (Mohammadi et al. 2018). U kontekstu HEMS-a, broj mogućih stanja i akcija može biti izuzetno veliki, zbog prisustva različitih tipova uređaja, vremenskih uslova, tarifnih režima i preferencija korisnika. Svaka dodatna dimenzija u prostoru stanja ili akcija eksponencijalno povećava veličinu Q-tabele. U realnim HEMS primenama, Q-tabela brzo postaje prevelika za efikasno čuvanje i ažuriranje, čime se značajno smanjuje njena praktičnost. Zbog toga se u razvoju HEMS-a sve više primenjuju modeli dubokog učenja potkrepljivanjem (eng. *Deep Reinforcement Learning*, DRL), koji omogućavaju efikasnije upravljanje velikim i složenim prostorima stanja.

Lissa i saradnici (Lissa et al. 2021) predložili su HEMS rešenje zasnovano na DQN algoritmu, sa ciljem optimizacije potrošnje električne energije korišćenjem proizvodnje iz solarnih panela. Fokus istraživanja bio je na smanjenju troškova električne energije povezanih sa grejanjem prostora i temperaturom tople vode. Definisana je funkcija nagrade sastavljena od pojedinačnih komponenti, koje imaju za cilj očuvanje komfora

stanovnika kroz kontrolu unutrašnje temperature prostora, temperature vode i korišćenje energije proizvedene putem solarnog panela:

- Komponente funkcije nagrade vezane za temperaturu imale su vrednost 0 kada je temperatura bila unutar zadatog opsega, dok se u suprotnom računala razlika između trenutne i srednje vrednosti temperature.
- Komponenta funkcije nagrade vezana za solarni panel imala je vrednost 0 ukoliko je energija proizvedena putem solarnog panela bila veća od ukupne energije potrebne za rad svih uređaja. U suprotnom, nagrada za solarni panel predstavljala je preostali iznos energije potreban za napajanje svih uređaja.

Konačna nagrada izračunavana je kao ponderisana suma pojedinačnih komponenti. Vrednost težinskog koeficijenta u funkciji nagrade određena je na osnovu analize zavisnosti između ukupne potrošnje energije i odstupanja temperature od zadatih granica, pri čemu je kao optimalna izabrana tačka neposredno pre one u kojoj dalji pad potrošnje uzrokuje značajno narušavanje komfora. Autori nisu razmatrali tarifne režime električne mreže niti su automatizovali upravljanje uređajima koji nisu povezani sa grejanjem, što bi moglo doprineti sveobuhvatnijem pristupu upravljanju energijom.

U radu (Forootani et al. 2022), problem više-ciljne optimizacije HEMS-a rešavan je dekompozicijom kontrole kompletnog skupa uređaja na pojedinačnu kontrolu svakog uređaja. Za svaki uređaj razvijen je nezavistan DQN model koji predviđa optimalne diskretne akcije specifične za taj uređaj. Modeli su obučavani odvojeno, sa ciljem minimizacije troškova električne energije i maksimizacije komfor stanara. Funkcija nagrade definisana je kao ponderisana suma dve komponente: (1) potrošnje električne energije i (2) komfora izraženog u vidu razlike između stvarnog i željenog vremena rada uređaja. Autori nisu ponudili metod za automatsku kalibraciju težinskih koeficijenata u funkciji nagrade, već su njihovo podešavanje prepustili stanarima.

Glavni nedostatak DQN algoritma jeste njegova ograničena efikasnost pri rešavanju problema koji se odlikuju velikim i kontinuiranim prostorima stanja (eng. *continuous state space*) (Mnih et al. 2015). U kontekstu HEMS-a, kontinuirani prostor stanja obuhvata varijable kao što su količina potrošene energije, količina energije proizvedene putem solarnih panela i vrednost temperature. Efikasnost treniranja DQN modela u ovakvom prostoru može se poboljšati primenom DDQN pristupa (Fu et al. 2022).

U radu (Liu et al. 2020) analizirana je efikasnost DQN i DDQN pristupa u obuci HEMS-a, čiji je cilj minimizacija troškova električne energije uz istovremeno očuvanje komfora stanovnika. Komfor korisnika kvantifikovan je vremenskim ograničenjima rada uređaja, pri čemu je funkcija nagrade uključivala negativni penal ukoliko uređaj nije aktiviran unutar željenog vremenskog opsega. Željeni opsezi rada su navedeni na osnovu istorijskih statistika i autori nisu dali smernice oko određivanja vrednosti penala. Autori su u svoju postavku problema uključili proizvodnju električne energije putem solarnih panela, mogućnost skladištenja energije u baterijama i tarifne režime električne mreže. Rezultati su pokazali da DDQN model postiže bolje performanse u odnosu na DQN model. U skladu sa tim nalazima, u ovom istraživanju se koristi DDQN pristup za obuku HEMS-a, sa ciljem rešavanja problema sličnog onom analiziranom u pomenutom radu.

Huang i saradnici (Huang et al. 2022) nastojali su da minimizuju operativne troškove sistema za upravljanje energijom u domaćinstvu, uz očuvanje termalnog komfora i poštovanje ograničenja vremena rada uređaja. Unutar domaćinstva analizirano je

upravljanje HVAC sistemom, obnovljivim izvorima energije, skladištem energije i uređajima sa ograničenim vremenom rada. HVAC sistem, koji obezbeđuje termalni komfor, kao i sistem za skladištenje energije, zahtevali su akcije koje podrazumevaju određivanje količine potrošene energije za rad uređaja ili, u slučaju baterija, količine energije za pražnjenje i punjenje. Zbog toga se za kontrolu ovih uređaja koristio DDPG pristup, čija je ciljna funkcija bila ponderisana suma potrošnje energije i termalnog komfora, definisanog kao razlika između postignute i željene temperature. Uređaji sa ograničenim vremenom rada zahtevali su diskretne akcije uključivanja i isključivanja. Stoga se za njihovu kontrolu koristio DQN pristup, gde je ciljna funkcija predstavljena kao ponderisana suma potrošnje energije i komfora u vezi sa preferiranim vremenom rada, izraženog kao razlika između stvarnog vremena rada uređaja i definisanog vremenskog okvira. Integracijom DQN i DDPG modela omogućeno je efikasno upravljanje kombinovanim diskretnim i kontinuiranim prostorom akcija u energetskom upravljanju pametnim kućama. Međutim, autori nisu naveli smernice za odabir težinskih koeficijenata u funkcijama nagrade.

Primena PPO modela mogla bi biti efikasnije rešenje u odnosu na treniranje dva odvojena modela, budući da PPO model podržava i diskretne i kontinuirane akcije, čime se smanjuje složenost sistema i olakšava optimizacija (Shakya et al. 2023).

Li i saradnici (Li et al. 2020) koristili su PPO algoritam za upravljanje HEMS-om sa tri tipa uređaja:

- Kritični uređaji, poput televizora, frižidera i osvetljenja, radili su na zahtev stanovnika i nisu bili pod kontrolom HEMS-a.
- Prilagodljivi uređaji, kao što su mašina za pranje sudova, veš mašina, mašina za sušenje i šporet, mogli su da prilagode vreme rada u skladu sa tarifnim režimima. HEMS je upravljao ovim uređajima koristeći diskretne akcije uključivanja i isključivanja.
- Kontrolisani uređaji, poput električnog vozila, imali su mogućnost fleksibilnog punjenja. HEMS je upravljao njihovom potrošnjom pomoću kontinuiranih akcija, određujući nivo snage za punjenje i pražnjenje baterije.

Cilj istraživanja bio je minimizacija troškova električne energije domaćinstva, uzimajući u obzir nesigurnost u cenama električne energije u realnom vremenu i neizvesnost u obrascima ponašanja stanara. Funkcija nagrade definisana je kao ponderisana suma dva faktora: (1) ukupne cene energije utrošene za rad uređaja i (2) kvadrata energije potrebne da se električno vozilo potpuno napuni. Autori nisu naveli kriterijume na osnovu kojih su određene težine ovih faktora u funkciji nagrade.

Koristeći opisanu postavku, *Li* i saradnici trenirali su PPO model i uporedili njegove performanse sa modelima zasnovanim na ručno definisanim pravilima – referentnim modelom i teorijski optimalnim modelom:

- Referentni model pokretao je sve uređaje (prilagodljive i kontrolisane) čim su oni bili spremni za rad, bez obzira na cenu električne energije ili zadati željeni opseg rada uređaja.
- Teoretski optimalan model predstavljao je ručno definisan raspored rada uređaja, usklađen sa unapred poznatim cenama električne energije. Ovaj model nije moguće implementirati u praksi jer podrazumeva poznavanje buduće cene električne energije, što je u ovoj postavci bila informacija koja je nedostupna

agentu u trenutku donošenja odluke. Zbog toga je ovaj model služio samo za procenu gornje granice performansi algoritma.

Rezultati su pokazali da je PPO model postigao značajno smanjenje troškova električne energije u odnosu na referentni model, dok su njegove performanse bile bliske performansama teoretski optimalnog modela.

Mbuwir i saradnici (Mbuwir et al. 2021) predstavili su hibridni pristup za punjenje električnih vozila, koji je kombinovao PPO algoritam sa sistemom kontrole zasnovanim na pravilima (eng. *rule-based*). Funkcija nagrade definiše se kao negativna suma troškova punjenja električnih vozila iz mreže i prihoda od prodaje viška energije nazad u mrežu. Takođe, u funkciju nagrade bio je uvršćen i faktor kazne, koji se primenjivao kada električno vozilo nije bilo u potpunosti napunjeno do heuristički procenjenog vremena polaska. Skup akcija bio je kontinuiran i predstavljao količinu snage (u *kW*) koja se koristila za punjenje električnog vozila. PPO model određivao je ovu akciju, dok je sistem zasnovan na pravilima služio kao kontroler koji je prilagođavao akciju predloženu od strane PPO modela, uzimajući u obzir trenutnu proizvodnju energije iz solarnog panela.

Mbuwir i saradnici su uporedili su predloženu kombinaciju PPO modela i ručno definisanih pravila sa dve strategije zasnovane na ručno definisanim pravilima:

- Naivna strategija, podrazumevala je punjenje vozila odmah po njegovom priključenju na stanicu za punjenje.
- Strategija sa perfektnim informacijama pretpostavljala je potpuno poznavanje svih relevantnih faktora dan unapred, uključujući vreme dolaska i odlaska električnog vozila, energiju potrebnu za punjenje, maksimalnu dozvoljenu snagu punjenja, količinu energije proizvedene putem solarnog panela i cenu električne energije. Ova strategija predstavljala je idealizovan scenario u kojem su sve informacije dostupne u trenutku donošenja odluke, dok su u realnim uslovima mnoge od tih informacija nepoznate u trenutku donošenja odluke.

PPO model sa pravilima je značajno smanjio troškove električne energije za 62.5% u poređenju sa naivnom strategijom. U poređenju sa strategijama sa perfektnim informacijama, ostvaruje rezultat koji je slabiji za samo 5%.

Sun i saradnici (Sun et al. 2020) razvili su više-agentni PPO algoritam za optimizaciju troškova električne energije u HEMS-u velikih razmera (eng. *large-scale*). Simulaciono okruženje obuhvatalo je više domaćinstava, pri čemu je svakom pametnom kućom upravljao poseban, nezavisni agent. Stanje pojedinačne kuće obuhvatalo je trenutnu cenu električne energije, minimalni nivo neophodne potrošnje za rad uređaja, kao i informacije o stanju električnog vozila, baterije i solarnog panela. Skup akcija bio je kontinuiran i uključivao je brzinu punjenja i pražnjenja električnih vozila i kućnih baterija, kao i donošenje odluka o trgovini električnom energijom unutar mikro-mreže (eng. *microgrid*). Cilj je bilo smanjenje troškova električne energije, pri čemu je cena bila dinamički određivana na osnovu odnosa ponude i potražnje u mreži. Funkcija nagrade definisana na nivou domaćinstva uključivala je: (1) trošak električne energije, (2) penalizaciju za potrošnju tokom perioda visoke tarife i (3) kolektivnu entropiju ponašanja (eng. *Collective Behavior Entropy*), koja je uvedena radi obeshrabrivanja sinhronizovanog opterećenja mreže.

Azuatalam i saradnici (Azuatalam et al. 2020) primenili su PPO algoritam za optimizaciju kontrole HVAC sistema u zgradi, sa ciljem balansiranja između očuvanja termalnog

komfora i smanjenja troškova električne energije. Skup mogućih akcija bio je kontinuiran i predstavljao je precizna podešavanja unutrašnje temperature. Simulaciono okruženje je bilo definisano uz pomoć *EnergyPlus*⁴ platforme, koja je simulirala HVAC sistem u zgradi. Kako bi se postigla realističnost simulacije, u uključeni su istorijski podaci o spoljnim vremenskim uslovima i solarnom zračenju. Funkcija nagrada definisana je kao ponderisana suma troškova električne energije i odstupanje unutrašnje temperature od željenog temperaturnog opsega. Težinski koeficijenti u funkciji nagrade bili su određeni empirijski, pri čemu autori nisu naveli smernice za njihovo podešavanje. U radu su upoređena tri pristupa:

- Osnovni model zgrade, koji je predstavljao jednostavne, ručno definisane strategije za upravljanje HVAC sistemom.
- Standardni PPO model.
- Kombinacija PPO modela sa dva dodatna kontrolera, koji su prilagođavala njegove akcije:
 - *Downward Demand Response* (DDR) kontroler, koji je smanjivao potrošnju energije tokom perioda visokih tarifa ili kada je bilo potrebno smanjiti opterećenja na mreži.
 - *Upward Demand Response* (UDR) kontroler, koji je povećavao potrošnju energije tokom perioda niskih cena električne energije ili kada je bilo potrebno povećanje opterećenja na mreži.

Najbolji rezultati postignuti su kombinovanjem PPO modela sa DDR i UDR kontrolerima. Konkretno, kombinacija PPO modela sa DDR kontrolerom pokazala je najbolje rezultate u smanjenju potrošnje energije tokom perioda visokih tarifa, dok je kombinacija PPO modela sa UDR kontrolerom bila najefikasnija u povećanju potrošnje energije tokom perioda niskih tarifa.

Aldahmashi i *Ma* (Aldahmashi and Ma 2024) definisali su novi optimizacioni problem koji integriše upravljanje aktivnom i reaktivnom električnom energijom u pametnim kućama, uz uvažavanje komfora stanara. Funkcija nagrade formulisana je kao ponderisana suma sledećih komponenti, pri čemu autori nisu naveli kriterijume za izbor težinskih koeficijenata:

- Ukupne električne energije potrebne za rad svih uređaja – električnog vozila, baterije i solarnog panela.
- Koeficijenta nezadovoljstva stanara, koji obuhvata odstupanje unutrašnje temperature od željenog opsega, kašnjenje u izvršavanju zadataka uređaja sa vremenskim ograničenjima, kao i nedovoljnu napunjenost baterije električnog vozila pre odlaska iz kuće.
- Penala koji se primenjivao u slučaju da faktor snage (eng. *power factor*) padne ispod zadatog minimalnog praga⁵, pri čemu je svakom uređaju bio dodeljen unapred definisan prag koji opisuje njegov odnos između potrošnje aktivne i reaktivne snage.

⁴ <https://energyplus.net/>

⁵ Faktor snage predstavlja odnos između aktivne snage (koja obavlja korisni rad) i prividne snage (ukupne snage u sistemu). Nizak faktor snage ukazuje na nisku energetske efikasnost, jer veći deo prenesene energije ostaje neiskorišćen za rad.

Autori su trenirali PPO model i uporedili ga sa sledećim alternativama:

- DQN i DDPG modelima, koji su trenirani korišćenjem iste funkcije nagrade, uz neophodne izmene skupova akcija i stanja radi prilagođavanja ovim algoritmima. Kod DQN-a, sve akcije, uključujući i kontinualne (na primer, snaga punjenja, reaktivna snaga), bile su diskretizovane zbog ograničenja DQN algoritma na diskretne prostore akcija. Kod DDPG-a, binarne akcije (na primer, uključivanje uređaja) transformisane su u kontinualni oblik, a zatim aproksimirane prilikom izbora konkretne akcije, kako bi bile kompatibilne sa prirodom kontinualnog akcijskog prostora koji DDPG zahteva.
- FIO (eng. *full information observable*) modelom, koji je imao pristup potpunim informacijama o svim relevantnim parametrima. U FIO modelu, dinamički promenljivi i unapred nepoznati faktori, poput proizvodnje električne energije putem solarnog panela, tretirani su kao deterministički i unapred poznati. Cilj uključivanja idealizovanog FIO modela bio je određivanje najefikasnije strategije planiranja vremena rada uređaja, radi procene gornje granice performansi treniranih modela. U okviru FIO modela, reaktivna energija nije bila optimizovana.

U izvedenim eksperimentima, PPO model je ostvario najbolje rezultate među metodama koje funkcionišu u uslovima nesigurnosti, zahvaljujući svojoj sposobnosti da u realnom vremenu optimizuje upravljanje aktivnom električnom energijom u bateriji i električnom vozilu. Postignuta ušteda iznosila je 31.5%, što je blisko maksimalnoj uštedi od 39.6% postignutoj idealizovanim FIO pristupom. Za poređenje, DDPG i DQN postigli su uštede od 24.8% i 18.6%, tim redosledom.

Na osnovu pregleda literature u ovom poglavlju, mogu se izvesti sledeći zaključci:

- Najperspektivnija rešenja za razvoj HEMS-a zasnovana na učenju potkrepljivanjem su DDQN i PPO algoritmi.
- Tipovi uređaja razmatrani u simulacijama pametnih kuća obuhvataju:
 - Uređaje za regulaciju temperature i uređaje sa vremenskim ograničenjem rada, koje HEMS kontroliše diskretnim akcijama uključivanja i isključivanja.
 - Skladišta energije, koja HEMS kontroliše kontinuiranim akcijama zadavanja snage punjenja ili pražnjenja uređaja.
 - Uređaje koji nisu regulisani od strane HEMS-a, uključujući uređaje koji rade isključivo na zahtev stanovnika (na primer, televizor) i proizvođače energije (na primer, solarni panel).
- Većina simuliranih pametnih kuća modeluje potrošnju uređaja kao konstantnu tokom rada iako uređaji poput veš mašine imaju složene obrasce potrošnje, sa složenim razlikama u potrošnji energije u različitim radnim fazama.
- Najčešće razmatrani aspekti komfora su termalni komfor i ograničenje vremena rada uređaja. Za kvantifikaciju komfora koriste se pristupi zasnovani na penalima i razlikama između definisanih i trenutnih vrednosti temperature, odnosno vremena rada uređaja.
- U većini slučajeva, više-ciljna optimizacija komfora i potrošnja električne energije realizuje se kroz ponderisane sume unutar funkcije nagrade. Pojedini pristupi dodatno uvode restrikcije rada uređaja i penale za nepoželjna ponašanja. Težinski koeficijenti i vrednosti penala najčešće se određuju empirijski, bez detaljnog metodološkog obrazloženja.

- Većina radova u okviru simulacije uzima u obzir tarifne režime i spoljašnje vremenske uslove.
- Većina radova koji predlažu pristupe zasnovane na učenju potkrepljivanjem poredi ove metode sa ručno definisanim strategijama, koje se koriste kao referentni modeli. Takve strategije su obično pojednostavljene, sa unapred poznatim obrascima potrošnje uređaja ili sa pojednostavljenim operativnim kriterijumima, što može ograničiti njihovu primenljivost u realnim uslovima.

3.3 Primena velikih jezičkih modela u sistemima za upravljanje energijom u pametnim kućama

Analiza literature pokazala je da su postojeće studije istraživale primenu LLM-ova u HEMS-u u tri glavna pravca:

- Simulacija okruženja pametne kuće,
- Razvoj pomoćnih alata za poboljšanje komfora korisnika,
- Podsticanje stanovnika na racionalnu potrošnju električne energije kroz formiranje energetske efikasne navike.

U radu (Almashor and Miyashita 2024), LLM-ovi su primenjeni za simulaciju okruženja pametne kuće. U razvijenoj simulaciji, agenti vođeni LLM-ovima imitirali su dnevne rutine i aktivnosti stanovnika koje su uticale na potrošnju energije, generišući obrasce potrošnje koji su oponašali stvarnu potrošnju energije u domaćinstvu. Simulacija je omogućavala kreiranje realističnih i javno dostupnih skupova podataka koji odražavaju energetske ponašanje specifičnih geografskih regiona. Ovim pristupom eliminisana je potreba za korišćenjem stvarnih, osetljivih podataka, čime je obezbeđeno očuvanje privatnosti. Međutim, glavni nedostatak ovog pristupa bila je pretpostavka da uređaji tokom rada troše konstantnu količinu energije. Uređaji poput veš mašine imaju složene obrasce potrošnje, sa značajnim razlikama u potrošnji energije tokom različitih radnih faza. Zbog toga je njihovo modelovanje kao uređaja sa konstantnom potrošnjom moglo rezultovati manje preciznim podacima o varijacijama u energetske potrošnji.

Giudici i saradnici (Giudici et al. 2024) istraživali su kako LLM-ovi mogu biti iskorišćeni za kreiranje rutina za automatizaciju domaćinstva koje podstiču ekološki svesno ponašanje. U njihovom radu, model GPT-4 (Koubaa 2023) korišćen je za asistenciju korisnicima pri kreiranju i kontroli automatizovanih rutina koje optimizuju potrošnju energije i promovišu održive navike. U interakciji sa korisnikom, GPT-4 imao je ulogu edukativnog agenta koji je vodio korisnike kroz proces kreiranja i prilagođavanja njihovih rutina. Glavno ograničenje ove studije bilo je to što je evaluirana na hipotetičkom scenariju, gde su korišćeni simulirani podaci umesto stvarnih interakcija sa kućnim uređajima. Alat nije pružao kompletnu automatizaciju, već je imao savetodavnu ulogu, prepuštajući donošenje odluka korisniku. Konačno, autori nisu evaluirali efikasnost sistema u redukciji potrošnje energije.

King i saradnici (King et al. 2024) primenili su LLM za interpretaciju i izvršavanje ciljano orijentisanih komandi koje su neprecizno formulisane, poput „učini prostor prijatnim“. Ovaj pristup omogućavao je korisnicima da izraze svoje želje na prirodni način, čime se poboljšavao njihov komfor. Na osnovu korisničke komande, LLM je generisao akcioni plan u formi JSON šablona, koji je detaljno opisivao kako se raspoloživi uređaji i senzori

mogu koristiti za postizanje željene atmosfere. Glavni nedostatak ovog pristupa bila je sklonost LLM-a da generiše nerelevantne ili čak potpuno netačne akcijske planove.

U radu (Oprea and Bâra 2024) predstavljen je sistem koji je vlasnicima kuća pružao savete za optimizaciju energetske transakcije i prilagođavanje potrošnje opterećenju lokalnog energetskeg tržišta. Sistem je koristio pretrenirani LLM kao klasifikator koji je primao tekstualne opise stanja okruženja i, na osnovu njih, predviđao optimalne akcije. Moguće akcije uključivale su kupovinu ili prodaju električne energije, kao i smanjenje ili povećanje njene potrošnje. Autori nisu analizirali uticaj određenih podataka iz tekstualnog opisa stanja okruženja na konačne predikcije klasifikatora.

Pregledom literature je utvrđeno da nijedna studija nije istraživala efikasnost pretreniranih LLM-ova u orkestraciji uređaja u pametnim kućama s ciljem optimizacije višeciljnog problema smanjenja troškova električne energije uz održanje komfora stanovnika. Stoga, postoji potreba za istraživanjem koje će ispitati mogućnost primene LLM-ova za rešavanje ovog problema.

3.4 Ograničenja postojećih rešenja i obrazloženje potrebe za daljim istraživanjem

Sistemi za upravljanje energijom u domaćinstvima predstavljaju izazovnu i dinamičnu oblast u kojoj još nisu standardizovani protokoli za optimizaciju i implementaciju rešenja (Mahapatra and Nayyar 2022). Pregledom literature identifikovane su značajne razlike u definisanju problema optimizacije, što otežava poređenje različitih rešenja. Ove razlike uključuju (Gomes et al. 2022) (Mahapatra and Nayyar 2022):

- **Cilj optimizacije:** različite studije koristile su različite funkcije nagrade za minimizaciju potrošnje električne energije uz očuvanje komfora stanovnika. Mere komfora bile su definisane na različite načine, a razlikovao se i način njihove integracije sa merama potrošnje energije u funkciji nagrade.
- **Okruženja:** simulaciona okruženja za pametne kuće razlikovala su se po dostupnim uređajima, njihovim karakteristikama i tarifnim režimima. *Matlab* je bio najčešće korišćen kao simulaciono okruženje, a nedostatak standardizovanog simulacionog okruženja istaknut je kao značajna prepreka za efikasno poređenje različitih rešenja (Gomes et al. 2022). Glavni nedostatak većine posmatranih okruženja odnosio se na nemogućnost modelovanja uređaja sa varijabilnom potrošnjom električne energije tokom različitih faza rada. U većini radova, uređaji poput veš mašine i mašine za sudove bili su modelovani tako da imaju konstantnu potrošnju tokom celokupnog perioda rada, što ne odgovara stvarnim karakteristikama njihove potrošnje.
- **Trening i test podaci:** modeli predloženi u analiziranim studijama obično su trenirani i testirani samo u određenim periodima godine, što je dovodilo u pitanje njihovu sposobnost generalizacije na druge periode sa različitim vremenskim uslovima i obrascima potrošnje. Da bi se ovo ograničenje prevazišlo, potrebno je obuhvatiti treniranje i testiranje modela tokom različitih sezona u godini.
- **Tehnike optimizacije ciljne funkcije:** Savremena istraživanja pretežno su se fokusirala na primenu tehnika učenja potkrepljivanjem za automatizaciju rada uređaja u pametnim kućama, sa ciljem minimizacije troškova energije i očuvanja komfora stanovnika. Međutim, nije pronađeno nijedno rešenje koje je razmatralo primenu pretreniranih LLM-ova za rešavanje ovog problema.

Takođe, utvrđeno je da postojeća rešenja nisu analizirala uticaj parametara stanja okruženja na odluke modela učenja potkrepljivanjem. Ovakva analiza mogla bi doprineti objašnjenju odluka donetih od strane HEMS-a, povećati poverenje korisnika u sistem i omogućiti njegovo rafiniranje.

Iz pregleda aktuelnog stanja u oblasti ustanovljeno je da problem optimizacije više-ciljne funkcije, koja uključuje minimizaciju troškova električne energije i održanje komfora stanara, nije u potpunosti rešen. Stoga postoji potreba za daljim istraživanjem u ovoj oblasti. U ovom istraživanju, navedeni problem biće rešavan automatizacijom planiranja vremena rada uređaja u pametnoj kući. Analizom literature utvrđeno je da rešenje treba da obuhvati sledeće aspekte:

- Definiciju adaptivnog okruženja za simulaciju pametne kuće. Okruženje bi trebalo da omogući simulaciju različitih tipova uređaja i fleksibilno podešavanje njihovih karakteristika. Kako bi simulacija uređaja bila što realističnija, poželjno je omogućiti simulaciju uređaja sa varijabilnom potrošnjom energije. Simulacija bi takođe trebala da podrži definisanje dostupnih tarifnih režima. Konačno, radi postizanja veće realističnosti eksperimenta, prilikom evaluacije rešenja je u simulaciju potrebno uključiti istorijske podatke o vremenskim uslovima i tarifnim režimima na određenoj geografskoj lokaciji.
- Definisanje metrika koje kombinuje komfor stanovnika pametne kuće sa cenom električne energije. Metrika komfora treba da omogući stanarima intuitivno definisanje željenog stepena i opseg rada pojedinačnih uređaja. Ove metrike se formalizuju putem funkcije nagrade, koja se koristi za obučavanje modela zasnovanih na učenju potkrepljivanjem, donošenje odluka od strane LLM-a, kao i za evaluaciju performansi RL i LLM modela.
- Odabir odgovarajućeg modela za treniranje HEMS-a. Potrebno je razmotriti DDQN i PPO algoritme učenja potkrepljivanjem, koji su prepoznati kao najperspektivnije strategije za rešavanje ovog problema. Takođe, neophodno je istražiti mogućnost primene LLM-ova, koji imaju potencijal da ublaže ograničenja navedenih algoritama. Kao referentni pristup, potrebno je definisati sistem zasnovan na ručno definisanim pravilima.
- Postupak evaluacije na trening i test podacima koji pokrivaju različita godišnja doba.
- Analizu uticaja ulaznih parametara stanja okruženja na odluke razvijenog HEMS sistema. Ovo istraživanje predlaže primenu SHAP (eng. *Shapley Additive Explanations*) biblioteke (Lundberg and Lee 2017), koja omogućava dekompoziciju predikcija modela na pojedinačne doprinose svakog ulaznog parametra.

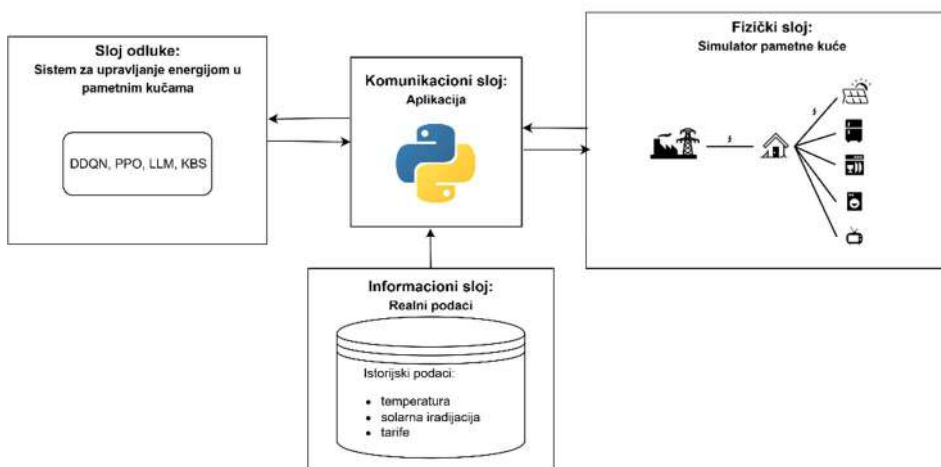
4 Metodologija

Ovo poglavlje opisuje metodologiju za optimizacije potrošnje električne energije uz očuvanje komfora stanovnika pametne kuće, ostvarenu strateškim planiranjem vremena rada uređaja. Potpoglavljje 4.1 izlaže arhitekturu implementiranog sistema. Potpoglavljje 4.2 opisuje dizajn adaptivnog simulacionog okruženja, dok potpoglavljje 4.3 opisuje funkciju nagrade. Potpoglavljje 4.4 opisuje primenu algoritama učenja potkrepljivanjem za rešavanje ovog problema. Potpoglavljje 4.5 objašnjava način korišćenja pretreniranih velikih jezičkih modela u ovom kontekstu, sa fokusom na dizajn efektivnih promptova.

4.1 Arhitektura sistema

U okviru ove disertacije, sistem je podeljen na četiri sloja, po uzoru na rad *Cook* i saradnika (Cook et al. 2003). Slika 18 prikazuje arhitekturu sistema raščlanjenu po slojevima:

- Simulaciono okruženje pametne kuće (fizički sloj) – Simulira okruženje i generiše podatke o potrošnji električne energije kućnih aparata. Radi evaluacije predloženog pristupa, kreirana je simulacija konkretne pametne kuće, čije su postavke opisane u potpoglavljju 5.1.
- Sloj za donošenje odluka (HEMS) – Predstavlja model koji donosi odluke na osnovu trenutnog stanja okruženja. Unutar ovog sloja moguće je menjati model odlučivanja u cilju analize efekata različitih rešenja na potrošnju električne energije i komfor stanovnika.
- Informacioni sloj – Prikuplja podatke relevantne za simulaciju. U eksperimentima izvedenim u okviru ove disertacije, korišćeni su istorijski podaci o vremenskim uslovima i tarifnim režimima na određenoj geografskoj lokaciji. Ove informacije su neophodne za preciznu simulaciju rada određenih uređaja i računanje troškova električne energije.
- Komunikacioni sloj – Omogućava razmenu podataka između HEMS-a i simulacionog okruženja. Implementiran je kao *Python* aplikacija i omogućava da se simulacija potrošnje električne energije prilagodi realnim uslovima.



Slika 18. Arhitektura sistema – komponente simulacionog okruženja za obuku i evaluaciju HEMS-a za upravljanje uređajima u pametnoj kući. U okviru sloja odluke moguće je odabrati jedan od sledećih pristupa: algoritam dvostruke duboke Q-mreže (DDQN), algoritam optimizacije proksimalne politike (PPO), jednu od dve varijante modela zasnovanog na velikom jezičkom modelu (LLM) ili model zasnovan na znanju (KBS).

4.2 Dizajn adaptivnog simulacionog okruženja

Za obučavanje i evaluaciju modela dubokog učenja i sistema zasnovanog na znanju, implementirana su dva adaptivna simulaciona okruženja:

- *Typhoon HIL* (Typhoon HIL 2008), u okviru koga se potrošnja uređaja modeluje kao konstantna vrednost tokom celokupnog trajanja rada uređaja,
- *Load Flow* (Vojnović et al. 2022) (Vojnović et al. 2023b) (Vojnović et al. 2023a), u okviru koga se potrošnja uređaja modeluje kao varijabilna vrednost, zavisna od faze rada uređaja.

U okviru simulacije modelovani su proizvođači i potrošači električne energije. Rad proizvođača energije (solarnog panela) zavisi od eksternih vremenskih uslova. Svi potrošači energije predstavljaju uređaje sa promenljivim vremenom rada, odnosno, uređaje čiji rad može biti odložen ili zakazan unutar vremenskih opsega koje definišu korisnici.

Potrošači se dalje klasifikuju u termalne i netermalne uređaje. Rad termalnih uređaja (na primer, klima uređaj) zavisi od sledećih faktora:

- Eksterni vremenski uslovi, poput temperature i sunčeve iradijacije.
- Termodinamičke karakteristike kuće, koje se u simulaciji definišu putem dva parametra: toplotna provodnost i koeficijent iskorišćenja klima uređaja.
- Željeni temperaturni opseg, koji definišu korisnici, a koji se koristi za regulaciju unutrašnje temperature u cilju očuvanja komfora.

Za netermalne uređaje (na primer, veš mašina), osim vremenskog okvira u kojem korisnici očekuju da uređaj završi svoj rad, moguće je dodatno definisati da li uređaj

zahteva neprekidni ciklus rada. Ukoliko HEMS prekine rad uređaja sa neprekidnim ciklusom rada, taj uređaj mora biti ponovo pokrenut od početka, pri čemu se sva prethodno utrošena energija gubi. Ovakav ishod se smatra neefikasnim i treba ga izbeći adekvatnim planiranjem vremena rada uređaja.

Pored toga, za netermalne uređaje uveden je indikator spremnosti za rad, koji označava da li je uređaj spreman da započne radni ciklus. Na primer, veš mašina je spremna za rad samo ako je prethodno napunjena vešom. Nakon završetka radnog ciklusa, uređaj se više ne smatra spremnim za rad, dok se ne ispuni novi uslov spremnosti. U simulaciji je eksplicitno onemogućeno pokretanje uređaja koji nije spreman za rad.

Vremenski korak t u simulaciji postavljen je na 15 minuta, iako sistem omogućava definisanje i drugih vrednosti. U kontekstu HEMS-a, vremenski korak predstavlja osnovnu jedinicu vremena u kojoj sistem donosi odluke. Na početku svakog vremenskog koraka, HEMS očitava trenutno stanje okruženja i, na osnovu toga, određuje odgovarajuće akcije za upravljanje uređajima. Doneta odluka se primenjuje tokom celog trajanja tog koraka. Izbor trajanja vremenskog koraka direktno utiče na ponašanje sistema: kraći vremenski koraci omogućavaju finiju kontrolu i brže reagovanje na promene u okruženju, ali povećavaju složenost obuke i zahteve u pogledu obrade podataka. Nasuprot tome, duži vremenski koraci smanjuju računarske zahteve, ali mogu ograničiti sposobnost sistema da se pravovremeno prilagodi dinamičnim uslovima, naročito u slučaju naglih promena u proizvodnji ili potrošnji energije. U postojećim radovima, vremenski korak obično traje od nekoliko minuta do jednog sata, kako bi se omogućila jednostavna primena algoritma u realnom sistemu pametne kuće (Yu et al. 2020). Radovi (Yu et al. 2020), (Lissa et al. 2021) i (Lee and Choi 2019) koriste vremenski korak od 1h, dok je u ovom istraživanju odabrana vrednost od 15 minuta u skladu sa praksom u relevantnoj literaturi, uključujući radove (Azuatalam et al. 2020) i (Aldahmashi and Ma 2024).

4.3 Dizajn funkcije nagrade

Problem optimizacije razmatran u ovoj disertaciji predstavlja više-ciljnu optimizaciju, pri čemu je određivanje međuzavisnosti i interakcije između pojedinačnih ciljeva kompleksno (Moffaert and Nowe 2014). Funkcija nagrade formulisana je kao ponderisana suma različitih ciljeva izraženih u novčanim jedinicama, čime se omogućava njihovo objedinjavanje u jedinstvenu funkciju i olakšava postupak optimizacije. U ovom radu razmatraju se dva ključna cilja: očuvanje komfora stanara, i minimizacija troškova električne energije. Oba cilja su kvantifikovana u istim monetarnim terminima (euro/kWh), što omogućava njihovo direktno sabiranje.

Komfor stanara u vremenskom koraku t računa se kao suma komfora svih $N_{\text{upravljiiv}}$ uređaja kojima HEMS upravlja:

$$R_t^{\text{komfor}} = \sum_{i=1}^{N_{\text{upravljiiv}}} \left(E_t^i + \frac{1}{1000_{\{E_t^i=0\}}} \right) * B_t^i * Kazna_t^i, \quad (27)$$

gde:

- E_t^i predstavlja energiju koju potroši uređaj i u vremenskom koraku t .

- B_t^i predstavlja novčani značaj dodeljen radu uređaja i u vremenskom koraku t , prema vrednostima prikazanim u tabeli 14.
- Indikatorska funkcija $\frac{1}{1000}\{E_t^i=0\}$ ima vrednost $\frac{1}{1000}$ ako uređaj i nije radio u vremenskom koraku t , a u suprotnom ima vrednost 0. Ova konstanta uvedena je kako bi se kazna primenila i na uređaje koji nisu aktivni, čime se izbegava nula u izrazu i obezbeđuje minimalni doprinos vrednosti komfora u skladu sa minimalnom potrošnjom uređaja.
- $Kazna_t^i$ predstavlja penalizaciju za nepotrebnu aktivaciju ili deaktivaciju uređaja i u vremenskom koraku t . Na primer, nepotrebna aktivacija nastupa kada se klima uređaj uključi iako je temperatura već u željenom opsegu. Bez uvedene kazne, HEMS bi u ovom slučaju dobio nagradu za komfor, uprkos nepotrebnoj potrošnji energije. Sa druge strane, nepotrebna deaktivacija nastupa kada se klima uređaj isključi iako unutrašnja temperatura nije u željenom opsegu.

Kazna se razlikuje u zavisnosti od tipa uređaja (termalni i netermalni uređaji):

$$Kazna_t^i = \begin{cases} Kazna_{t,termalni}^i, & \text{ako je } i \text{ termalan uređaj,} \\ Kazna_{t,netermalni}^i & \text{ako je } i \text{ netermalan uređaj.} \end{cases} \quad (28)$$

Kazna za termalne uređaje može poprimiti tri različite vrednosti:

$$Kazna_{t,termalni}^i = \begin{cases} -1, & \Delta T_{t,razlika} = 0 \text{ i} \\ & A_{uključi}^i \in A_t \text{ i} \\ -|\Delta T_{t,razlika}| * 10, & \Delta T_{t-1,razlika} = 0, \\ & \Delta T_{t,razlika} \neq 0 \text{ i} \\ 1, & A_{uključi}^i \notin A_t, \\ & \text{inače.} \end{cases} \quad (29)$$

gde je:

- $\Delta T_{t,razlika}$ – razlika između trenutne unutrašnje temperature i željenog temperaturnog opsega u vremenskom koraku t ,
- $\Delta T_{t-1,razlika}$ – razlika u prethodnom vremenskom koraku $t - 1$,
- $A_{uključi}^i$ – akcija uključivanja termalnog uređaja i ,
- A_t – skup akcija preduzetih u vremenskom koraku t .

Na primer, kazna od -1 primenjuje se ako HEMS uključi klima uređaj iako je unutrašnja temperatura već bila unutar željenog opsega u prethodnom i trenutnom vremenskom koraku, čime se sprečava dodela nagrade za komfor kroz nepotrebnu potrošnju energije. S druge strane, ako je temperatura van željenog opsega, a HEMS ne aktivira AC, kazna se povećava proporcionalno odstupanju od željenog temperaturnog opsega. U suprotnom nema kazne ($Kazna_{t,termalni}^i = 1$).

Kazna za netermalne uređaje uvedena je da obeshrabri HEMS da bespotrebno isključuje uređaje:

$$Kazna_{t,\text{netermalni}}^i = \begin{cases} -1, & \text{broj ciklusa}_{t,\text{netermalni}}^i = 0 \text{ i } S_{t-1,W}^i = 0 \text{ i } A_{\text{isključ}}^i \in A_t, \\ 1, & \text{u suprotnom.} \end{cases} \quad (30)$$

Kazna od -1 primenjuje se u sledećem slučaju:

- uređaj je završio sve cikluse ($\text{broj ciklusa}_{t,\text{netermalni}}^i = 0$),
- uređaj je bio isključen u prethodnom koraku ($S_{t-1,W}^i = 0$),
- HEMS pokušava da isključi uređaj ($A_{\text{isključ}}^i \in A_t$), iako to nije potrebno.

Postoji dodatna kazna za netermalne uređaje (mašina za sudove i mašina za veš).

Neposredna nagrada na kraju dana $\tilde{R}_{\text{komfor}_T}$ računa se kao neposredna nagrada u vremenskom koraku T, uvećana za eventualne kazne koje se primenjuju ukoliko netermalni uređaji nisu završili svoj posao:

$$\tilde{R}_{\text{komfor}_T} = R_T + \sum_i^{N_{\text{netermalni}}} kazna_T^i. \quad (31)$$

Kazna $kazna_T$ uzima vrednost 0 ukoliko je uređaj i završio svoj rad do kraja dana (vremenski trenutak T). U suprotnom, predstavlja fiksnu novčanu penalizaciju koja je empirijski određena tako da bude približna najvećoj prosečnoj nagradi svih modela. Na osnovu eksperimentalne evaluacije (potpoglavljje 6.1), ukupna dnevna nagrada kreće se u intervalu od -6 do 4 (6.1.1), gde je najveća prosečna dnevna nagrada oko 4 evra za letnji mesec prilikom obučavanja PPO modela u *Typhoon HIL* simulacionom okruženju. Stoga je kazna postavljena na -4 evra po uređaju koji ne završi svoj rad do kraja dana, kako bi postojao značajan uticaj na ukupnu nagradu u slučaju neizvršavanja zahteva.

Ukupni trošak električne energije u vremenskom koraku t računa se kao suma potrošnje svih aktivnih uređaja:

$$R_t^{\text{potrošnja}} = \sum_{i=1}^N (E_t^i * \text{tarifa}_t), \quad (32)$$

gde:

- N predstavlja broj aktivnih uređaja, uključujući uređaje pod kontrolom HEMS-a (čiji je broj $N_{\text{upravljiv}}$) i uređaje koji predstavljaju proizvođače energije.
- E_t^i označava količinu energije koju i -ti uređaj potroši ili proizvede u vremenskom koraku t . Vrednost E_t^i je pozitivna ako uređaj troši energiju, a negativna ako je proizvodi.
- tarifa_t predstavlja cenu električne energije u vremenskom koraku t .

Neposredna nagrada (R_t) u vremenskom koraku t definiše se kao zbir komponente koja kvantifikuje komfor stanara za uređaje kojima upravlja HEMS (R_t^{komfor}) i komponente koja predstavlja trošak potrošnje električne energije svih aktivnih uređaja u tom vremenskom koraku ($R_t^{\text{potrošnja}}$):

$$R_t = R_t^{\text{komfor}} + R_t^{\text{potrošnja}}. \quad (33)$$

U finalnom vremenskom koraku $T = 96$, neposredna nagrada R_T se računa kao:

$$R_t = \tilde{R}^{\text{komfor}}_T + R^{\text{potrošnja}}_T,$$

gde se $R^{\text{potrošnja}}_T$ računa na osnovu (32), dok se $\tilde{R}^{\text{komfor}}_T$ računa na osnovu (31), koja je izmenjena nagrada za komfor koja podržava kazne za netermalne uređaje.

4.4 Dizajn modela učenja potkrepljivanjem za određivanje optimalnog vremena rada uređaja pametne kuće

Pregledom literature (potpoglavlje 3.2) utvrđeno je da su DDQN i PPO algoritmi najperspektivnija rešenja za razvoj HEMS-a zasnovanog na učenju potkrepljivanjem. Shodno tome, ovi algoritmi su primenjeni u ovom istraživanju.

U ovom istraživanju, rešavani problem modelovan je kao epizodni zadatak, pri čemu jedna epizoda odgovara jednom danu simulacije⁶. Tokom obučavanja modela učenja potkrepljivanjem, svaki naredni dan bira se nasumično, kako bi se izbegla zavisnost između uzastopnih epizoda. Na kraju svake epizode, simulacija se resetuje u početno stanje. Reset uključuje postavljanje unutrašnje temperature na vrednost $\max(18^\circ\text{C}, \text{Tempatura}_{0,\text{spoljašnja}})$. Takođe, indikatori spremnosti svih uređaja se resetuju i postavljaju u stanje „spremnost za rad“.

Postavke neuronskih mreža u DDQN i PPO algoritmima imaju identične dimenzije ulaznog i izlaznog sloja. Ulazni sloj ima deset dimenzija, što odgovara komponentama vektora stanja definisanim u potpoglavlju 5.1.4.1. Izlazni sloj ima sedam dimenzija i predstavlja ukupan broj mogućih akcija koje su opisane u potpoglavlju 5.1.4.2.

Slična postavka problema prisutna je u literaturi, pri čemu dužina epizode i inicijalni uslovi variraju u različitim istraživanjima:

- U radu Aldahmashi and Ma (Aldahmashi and Ma 2024), jedna epizoda trajala je jedan dan. Nakon svake epizode, okruženje se resetovalo tako što se unutrašnja temperatura nasumično postavljala u opsegu od 20°C do 22°C , dok su ostale varijable imaju unapred definisane početne vrednosti. Prilikom obuke PPO modela korišćeno je 300 nasumično sortiranih dana.
- U radovima Yu et al. (Yu et al. 2020) i Lee and Choi (Lee and Choi 2019), epizoda je takođe trajala jedan dan. Autori nisu eksplicitno naveli inicijalne vrednosti početnog stanja, već su definisali fiksne vrednosti za hiperparametre, potrošnju uređaja i minimamalnu/maksimalnu temperaturu.
- U radu Lissa et al. (Lissa et al. 2021), epizoda je trajala osam meseci (od maja do decembra). Autori su naveli samo vrednosti hiperparametara i željeni temperaturni opseg, ali nisu eksplicitno opisali inicijalno stanje.
- U radu Azuatalam et al. (Azuatalam et al. 2020), epizoda je trajala dva meseca (januar i februar). Nije naglašeno da se okruženje resetovalo između epizoda, ali su navedene inicijalne vrednosti hiperparametara i druge konstante relevantne za simulaciju.

⁶ Budući da se simulacija odvija u 15-minutnim koracima, jedna epizoda sadrži 96 vremenskih koraka.

- U radu Li, Wan i He (Li et al. 2020), epizoda je trajala jedan dan. Autori su eksplicitno naveli inicijalne vrednosti hiperparametara i ostale parametre neophodne vrednosti za simulaciju.

Radovi Azuatalam et al. (Azuatalam et al. 2020) i Aldahmashi and Ma (Aldahmashi and Ma 2024) pokazali su da PPO model brže konvergira u poređenju sa DDQN modelom. U radu (Azuatalam et al. 2020) PPO algoritmu bilo je potrebno oko 1000 epizoda za obuku. U njihovoj postavci, jedna epizoda trajala je dva meseca, uz vremenski korak od 15 minuta, što odgovara otprilike 5.760 vremenskih koraka po epizodi. Slično tome, u radu (Aldahmashi and Ma 2024), PPO model je konvergirao nakon približno 1.000 epizoda, dok je DDQN zahtevao oko 3.500 epizoda da bi postigao stabilnu konvergenciju. U njihovom eksperimentu, svi modeli su se trenirali 5.000 epizoda. Epizoda je trajala jedan dan i obuhvatala 96 vremenskih koraka (vremenski korak od 15 minuta). Li, Wan i He (Li et al. 2020) trenirali su PPO algoritam na 50.000 epizoda, gde je svaka epizoda trajala jedan dan simuliran u vremenskim koracima od 15 minuta. U ovoj tezi, PPO algoritam treniran je kroz 11.000 epizoda u okviru oba simulaciona okruženja, dok je za DDQN model korišćeno 22.000 epizoda u *Load Flow* okruženju i 44.000 epizoda u *Typhoon* HIL okruženju.

Potpoglavlja 4.4.1 i 4.4.2 prikazuju odabir ostalih hiperparametara DDQN i PPO algoritma.

4.4.1 Postavka hiperparametara za model dvostruke duboke Q-mreže

Odabir arhitekture duboke Q-mreže zavisi od dimenzije prostora stanja i akcija. Na primer, u radu (Aldahmashi and Ma 2024), prostor stanja bio je predstavljen kao vektor od 16 realnih vrednosti, dok je skup akcija obuhvatao 576 diskretnih vrednosti. Zbog visoke složenosti ove postavke, autori su dizajnirali DDQN model sa šest skrivenih slojeva – tri sloja sa po 128 neurona i tri sloja sa po 64 neurona. Kao aktivaciona funkcija korišćena je ReLU, a vrednost faktora umanjenja postavljena je na 0,99. DDQN model iz rada (Li et al. 2020) imao je jednostavniju arhitekturu, sa dva skrivena sloja od po 64 neurona. Ovakva konfiguracija odgovarala je manje složenoj postavci – prostor stanja sadržao je 33 realne vrednosti, ali je skup akcija bio znatno jednostavniji i sadržao svega pet mogućih akcija (četiri binarne i jednu kontinulanu). I u ovom slučaju korišćena je ReLU aktivaciona funkcija, uz faktor umanjenja od 0,995.

U ovom istraživanju, prostor stanja i akcija bio je jednostavniji u poređenju sa prethodno analiziranim radovima. Prostor stanja modelovan je kao 10-dimenzioni vektor koji uključuje realne i celobrojne vrednosti, dok skup akcija obuhvata sedam diskretnih mogućnosti (potpoglavlje 5.1.4). U skladu sa manjom složenošću problema, korišćena je potpuno povezana neuronska mreža sa dva skrivena sloja od po 64 neurona. Kao i u analiziranim radovima, korišćena je ReLU aktivaciona funkcija, dok je faktora umanjenja postavljen na 0,99.

Rad (Fuente and Guerra 2024) prikazao je eksperimentalnu postavku za DDQN i PPO modele, u okviru koje su upoređivane krive obučavanja za različite vrednosti stope učenja. Najvišu srednju vrednost nagrade postigla je stopa učenja $5e - 5$, dok je vrednost $5e - 4$ pokazala sličnu dinamiku konvergencije. U skladu sa ovim analizama, u okviru ovog istraživanja ispitano je više vrednosti stope učenja unutar tog raspona. Optimalna

vrednost odabrana je na osnovu vizuelne analize krivih obučavanja (primeri krivih obučavanja su prikazani na slikama u potpoglavlju 6.1.1.

Radi rešavanja kompromisa između istraživanja i eksploatacije, u ovom istraživanju primenjena je ϵ -pohlepna strategija sa linearnim opadanjem parametra ϵ , i opsegu od 1 do 0.1. Ovakva šema smanjenja podstiče intenzivno istraživanje u ranim fazama obuke, dok postepeno prelazi na eksploataciju naučene politike u kasnijim fazama treniranja. Linearno smanjenje ϵ omogućava postepen i stabilan prelazak između faza istraživanja i eksploatacije. I pored ograničenja poput nekonzistentnosti u izboru akcija tokom vremena (Plappert et al. 2018), ϵ -pohlepna strategija ostaje široko korišćena u praksi zbog jednostavne implementacije i robustnih performansi tokom procesa obuke.

Konačan izbor hiperparametara za DDQN model bio je prilagođen specifičnostima podataka generisanih korišćenjem različitih simulacionih okruženja. Vrednosti hiperparametara određene su na osnovu vizuelne analize krivih obučavanja (odnosno prosečne nagrade po epizodi), kao i u skladu sa preporukama iz relevantne literature. Tokom obuke, hiperparametari poput stope opadanja istraživanja (eng. *exploration rate decay*), stope učenja i ukupnog broja epoha varirali su u zavisnosti od konkretnih karakteristika simulacionog okruženja. Podaci dobijeni iz *Typhoon HIL* simulacionog okruženja zahtevali su veći broj epizoda za postizanje konvergencije. Budući da stope učenja i stope opadanja istraživanja zavise od ukupnog trajanja obuke, njihove optimalne vrednosti su se razlikovale u poređenju sa vrednostima korišćenim u *Load Flow* simulacionom okruženju. Optimalne vrednosti hiperparametara korišćenih za DDQN model prikazane su u tabeli 5.

Tabela 5. Vrednosti hiperparametara korišćene za obučavanje DDQN modela.

Hiperparametar	Vrednost	
	<i>Typhoon HIL</i>	<i>Load Flow</i>
Aktivaciona funkcija	ReLU	
Veličina skupa i memorijskog bafera	64/100.000	
Faktor umanjenja (γ)	0,99	
Broj vremenskih koraka po epizodi	96	
Broj epizoda	44.000	11.000
Stopa opadanja faktora istraživanja (ϵ)	$0,99/(44.000 * 96)$	$0,99/(11.000 * 96)$
Krajnji vrednost faktora istraživanja (ϵ_{min})	0,1	
Broj skrivenih slojeva	2 sloja od po 64 neurona	
Stopa učenja (α)	$2,5e-5$	$2,5e-4$
Funkcija greške	Huber	
Optimizator	Adam	

4.4.2 Postavka hiperparametara algoritma optimizacije proksimalne politike

U ovom istraživanju primenjena je ista postavka PPO algoritma kao u radu (Azuatalam et al. 2020), uz prilagođavanje vrednosti hiperparametara specifičnostima razvijenog simulacionog okruženja i ciljevima eksperimenta.

U navedenom radu, stanje sistema bilo je predstavljeno pomoću 15 kontinualnih vrednosti, dok je skup akcija sadržao 10 kontinualnih vrednosti. Mreže politike i vrednosti imale su po dva skrivena sloja sa po 64 neurona i *tanh* aktivacionom funkcijom. Parametar *clip* iznosio je 0,2, a faktor umanjenja $\gamma = 0,99$, čime se obezbeđuje stabilna ravnoteža između trenutnih i budućih nagrada.

U poređenju sa prethodnim istraživanjima, gde je broj epizoda često prelazio nekoliko desetina hiljada (npr. 50.000 u radu (Li et al. 2020)), u ovom istraživanju su vrednosti hiperparametara određene tako da obezbede kompromis između stabilnosti obuke, brzine konvergencije i vremenske složenosti simulacije.

Na osnovu rezultata radova radu (Azuatalam et al. 2020) i (Aldahmashi and Ma 2024), gde je PPO model konvergirao znatno brže od DDQN-a (oko 1.000 epizoda naspram 3.500), definisane su sledeće vrednosti: 11.000 epizoda u Load Flow okruženju, i 22.000 epizoda u Typhoon HIL okruženju. Svaka epizoda obuhvata 96 vremenskih koraka (odgovara jednom danu simulacije sa vremenskim korakom od 15 minuta). Ovim izborom omogućeno je postizanje stabilne konvergencije uz prihvatljivo trajanje simulacija i konzistentnost sa referentnim radovima.

Parametar *clip* = 0,2 preuzet je iz originalnog PPO rada (Schulman et al. 2017) i potvrđen kao stabilan u istraživanjima navedenih autora, dok su stope učenja odabrane eksperimentalno: 3×10^{-4} za mrežu politike i 1×10^{-3} za mrežu vrednosti. Ove vrednosti su se pokazale pogodnim za brzu stabilizaciju gradijenta bez oscilacija tokom treninga.

Broj slojeva i neurona (dva skrivena sloja sa po 64 neurona) preuzet je iz radova (Azuatalam et al. 2020) i (Li et al. 2020), čime je očuvana arhitektonska jednostavnost uz dovoljnu reprezentacionu moć modela.

Korišćene vrednosti hiperparametara PPO modela prikazane su u tabeli 6.

Tabela 6. Vrednosti hiperparametara korišćeni za obučavanje algoritma optimizacije proksimalne politike.

Hiperparametri	Vrednosti	
	<i>Typhoon HIL</i>	<i>Load Flow</i>
Aktivaciona funkcija	<i>tanh</i>	
<i>Clip</i> parametar	0,2	
Faktor umanjenja	0,99	
Broj vremenskih koraka po epizodi	96	
Broj epizoda	22.000	11.000
Broj skrivenih slojeva	2 sloja od po 64 neurona	
Stopa učenja (mreža politike)	3e-4	
Stopa učenja (mreža vrednosti)	1e-3	

4.5 Dizajn prompta koji omogućava velikom jezičkom modelu da odredi optimalno vreme rada uređaja pametne kuće

U ovom potpoglavlju prikazane su dve strategije definisanja prompta pri primeni LLM-a na problem automatskog određivanja vremena rada uređaja u cilju smanjenja potrošnje električne energije uz očuvanje komfora stanara.

potpoglavlje 4.5.1 prikazuje implementaciju strategije bez primera (eng. *zero-shot*), u kojoj je modelu data instrukcija da generiše kod koji predstavlja HEMS zasnovan na pravilima. Cilj ovog pristupa je provera da li LLM može automatski generisati sistem zasnovan na pravilima, na način sličan ljudskom ekspertu. Stoga je sistem zasnovan na pravilima generisan od strane LLM-a upoređen sa sistemom zasnovanim na pravilima koji je definisao autor disertacije.

U potpoglavlju 4.5.2 prikazana je *few-shot* strategija, u okviru koje je LLM tretiran kao model učenja potkrepljivanjem. Model je primao niz promptova, pri čemu je svaki prompt odgovarao jednom vremenskom koraku simulacije. U okviru svakog prompta, modelu je bio prosleđen opis trenutnog stanja, na osnovu koga je trebalo da odabere optimalnu akciju. Pored toga, model je imao pristup ograničenom broju prethodnih interakcija, koje su uključivale opise prethodnih stanja, odabranih akcija i pripadajuće vrednosti neposredne nagrade. Ovaj pristup omogućava imitaciju procesa donošenja odluka karakterističnog za učenje potkrepljivanjem, bez potrebe za dodatnim treniranjem modela.

Promptovi korišćeni u navedenim strategijama definisani su na osnovu tehnika predstavljenih u potpoglavlju 2.2.2.3. Ove tehnike pružaju strukturirani obrazac za formatiranje promptova, čime se obezbeđuje jasna interpretacija zadatka, doslednost u odgovorima i optimizacija akcija u skladu sa ciljevima HEMS-a. Ove dve strategije, uključujući KBS definisan od strane eksperta, se razlikuju od modela zasnovanih na učenju potkrepljivanjem jer kao rezultat mogu da vrate više akcija u vremenskom koraku.

4.5.1 *Zero-shot* strategija prompta

Zero-shot strategija koristi LLM za generisanje sistema baziranog na znanju, polazeći od prompta koji opisuje problem. Prompt je razvijan iterativno: u svakoj iteraciji LLM je dobijao opis problema uz zadatak da identifikuje nejasnoće i predloži dopune. Na osnovu konačne verzije prompta, od modela je zatraženo da generiše *Python* kod koji implementira sistem baziran na znanju. Tekst prompta je prikazan u tabeli 7.

Tabela 7. Tekst prompta na osnovu koga je LLM generisao *Python* kod koji implementira sistem zasnovan na znanju za upravljanje uređajima pametne kuće.

Prompt	Objašnjenje
You are a house energy management system. You must manage when appliances work so that electricity costs are reduced, while ensuring that the resident's comfort is maintained.	Definiše se osnovna funkcija sistema za upravljanje energijom u domaćinstvu – planiranje vremena rada uređaja radi smanjenja troškova električne energije, uz očuvanje komfora stanara. Naglašeno je da više uređaja može raditi istovremeno.

Multiple devices are allowed to run in parallel.	
The house has a solar panel and is connected to the electrical grid. The grid has the following fixed schedule for tariffs: peak (0.4 eur/kWh), shoulder (0.3 eur/kWh), and off-shoulder (0.2 eur/kWh).	Kuća je opremljena solarnim panelom i priključena na električnu mrežu. Navedene su fiksne tarife koje zavise od doba dana.
The house also has controllable devices: a washing machine, a dishwasher, and an AC. The energy consumption rates are variable throughout their run.	Navedeni su uređaji čijim radom sistem može upravljati: mašina za veš, mašina za sudove i klima uređaj. Ukazano je da je njihova potrošnja energije promenljiva tokom ciklusa rada.
The washing machine has a defined duration of 4 h 30 min, with consumption rates from 0 kWh to around 0.5 kWh. The dishwasher has a defined duration of 2 h, with consumption rates from 0 kWh to around 0.45 kWh. The washing machine and dishwasher can only be run when "available" (e.g., when loaded). Currently, both devices are ready at the start of the day. Once run, they cannot run again. These two appliances must run once daily; failure to run them results in a large, fixed penalty in the reward function.	Opisane su karakteristike mašine za veš i mašine za sudove, uključujući trajanje ciklusa i opseg potrošnje energije. Istaknuto je da ovi uređaji mogu raditi samo kada su spremni za rad i da moraju biti pokrenuti tačno jednom u toku dana.
The AC has a variable period, can be run multiple times per day, and has consumption rates ranging from 0 kWh to around 0.6 kWh.	Opisane su karakteristike klima uređaja. Istaknuto je da klima uređaj može biti uključen više puta dnevno.
The priorities of each device (washing machine, dishwasher, and AC) are different, time-dependent, and defined by users. For example, the washing machine's default priority is low, but if the residents want laundry done by 5 PM and it takes 2 hours, the priority becomes mid at 1 PM and high at 3 PM, signaling urgency.	Prioriteti uređaja definišu se od strane korisnika i vremenski su promenljivi. Definicija prioriteta se ilustruje kroz primer mašine za veš.
Residential comfort is modeled in monetary terms. Each controllable device at each timestep is assigned a priority value: high (0.45 eur/kWh), mid (0.35 eur/kWh), and low (0 eur/kWh).	Komfor se modeluje u novčanim jedinicama i zavisi od prioriteta uređaja.
The immediate reward function at each timestep is the sum of the benefit and the total energy	Vrednost neposredne nagrade u svakom vremenskom koraku predstavlja zbir komfora i potrošnje električne energije. Komfor pojedinačnog uređaja se računa kao proizvod potrošnje energije

consumption and production of each device (in euros).	tog uređaja, njegovog prioriteta i eventualne kazne vezane za rad uređaja.
Specifically, comfort is calculated as the sum of the product of energy consumption, benefit, and penalty for each device in the current timestep.	Ova komponenta komfora formalno je definisana sumom prikazanom u formuli (27), gde je izražena kao zbir doprinosa svih upravljivih uređaja u datom vremenskom koraku.
There is a penalty factor if the dishwasher or washing machine are stopped during operation.	Kazne vezane za rad uređaja se primenjuju u specifičnim situacijama: kažnjava se prekid ciklusa rada mašine za veš ili mašne za sudove, dok klima uređaj dobija kaznu ukoliko razlika između željene i stvarne unutrašnje temperature postane prevelika. Naglašava se da ukupna potrošnja energije može biti i negativna, što ukazuje da je proizvodnja energije putem solarnog panela veća od potrošnje neophodne za rad uređaja.
The AC incurs a penalty if the temperature difference between desired and actual indoor temperatures becomes too high.	
If total energy consumption is negative, this indicates excess production from the solar panel.	
The energy produced by the solar panel is directly used by devices, but surplus energy cannot be sold or stored.	Energija proizvedena solarnim panelom mora se odmah iskoristiti, jer se višak ne može ni skladištiti ni prodati. Količina proizvedene energije zavisi od vremenskih uslova, pri čemu se kao ulaz koriste realni istorijski meteorološki podaci.
Production depends on external weather conditions; real-world historical data is used as input.	
The final reward at the end of the day is calculated by summing the immediate rewards from each timestep.	Ukupna nagrada za jedan dan dobija se sabiranjem trenutnih nagrada iz svih vremenskih koraka. Vremenski koraci su diskretni i traju po 15 minuta. Svaka epizoda modeluje jedan nezavisan dan.
Timesteps occur in discrete intervals of 15 minutes.	
Each episode represents one independent day.	

Algoritam 4 definiše kriterijume za upravljanje uređajima na osnovu koda generisanog LLM-om. Algoritam 5 predstavlja sistem zasnovan na znanju koji je definisao autor disertacije. Oba algoritma, na osnovu stanja sistema (definisanim u (34)) i skupa mogućih akcija (definisanim u (35)), odlučuju koje uređaje treba uključiti ili isključiti u datom vremenskom koraku. Ulaz predstavlja trenutno stanje sistema, dok izlaz čini skup odabranih akcija.

Ulaz u algoritam 4 čini stanje sistema (S). Na izlazu algoritam vraća kolekciju akcija koje definišu koje uređaje treba uključiti u narednom vremenskom koraku.

Tokom izvršavanja, algoritam iterira kroz sve uređaje u sistemu i za svaki od njih određuje uslove za aktivaciju. Sa ovim ciljem se definišu pomoćne promenljive:

- *proizvedena_struja_prelazi_prag*, koja označava proveru da li proizvedena solarna energija u trenutnom vremenskom koraku prelazi prag od $0,1 kWh$. Vrednost praga od $0,1 kWh$ uvedena je od strane LLM-a, kako bi se izbegli šumovi u merenju i sprečilo uključivanje uređaja kada količina proizvedene

energije nije dovoljna za njihovu upotrebu. Prag je određen od strane LLM-a na osnovu minimalne potrebne energije za rad uređaja.

- *niska_tarifa*, koja označava da li je tarifa električne energije na najnižem nivou.

Ukoliko je reč o netermalnom uređaju (mašina za veš ili mašina za sudove), koji je spreman za rad, a nije radio tokom tekućeg dana, algoritam donosi odluku o njegovom uključivanju na osnovu značaja rada. Ako je značaj označen kao “Visok”, uređaj se uključuje. Ako je označen kao “Srednji”, uređaj se uključuje samo u slučaju da solarna proizvodnja prelazi prag ili da je tarifa “Niska” ili “Srednja”. Ako je značaj rada označen kao “Nije važno”, uređaj se uključuje samo ako je proizvodnja solarna energije veća od praga.

Aktivacija klima uređaja zavisi od njegovog značaja rada i uslova okruženja. Ako je značaj rada označen kao “Visok” ili “Srednji”, uređaj se uključuje u jednom od sledećih slučajeva:

- proizvodnja solarne energije prelazi zadati prag ili je trenutno niska tarifa.
- Razlika između željene i trenutne temperature je veća od 2 °C, a značaj rada je “Visok”. Vrednost od 2 °C definiše maksimalnu dozvoljenu temperaturnu razliku, pri čemu željeni temperaturni opseg, u rasponu od 20 °C do 25 °C, određuje korisnik. Toleranciju od 2 °C je LLM zadao na osnovu istraživanja (Buratti et al. 2015), gde je pokazano da temperaturna razlika veća od 2 °C izaziva izražen subjektivni osećaj nelagodnosti, odnosno osećaj da je prostor previše hladan ili topao.

Algoritam 4. Algoritam za određivanje koji uređaji treba da rade u datom vremenskom koraku, zasnovan na sistemu baziranom na znanju koji je generisao LLM. Oznake korišćene u algoritmu odgovaraju onima iz prikaza stanja *S* (jednačina (34)) i iz skupa akcija *A* (jednačina (35)):

- $S_{\Delta T_{razlika}}$ označava razliku između željene i trenutne unutrašnje temperature
- S_p^{AC} je značaj rada klima uređaja
- $(S_W^{CW}, S_L^{CW}, S_p^{CW})$ i $(S_W^{DW}, S_L^{DW}, S_p^{DW})$ odgovaraju mašini za veš i mašini za sudove i označavaju da li je uređaj već radio tokom dana, da li je uređaj spreman za rad i značaj rada uređaja.
- *uređaji* predstavlja skup svih uređaja.
- $A_{uređaj\ uključiti}$ označava akciju uključivanja uređaja.

Ulaz: Stanje $S = (S_{\Delta T_{razlika}}, S_p^{AC}, S_W^{CW}, S_L^{CW}, S_p^{CW}, S_W^{DW}, S_L^{DW}, S_p^{DW}, S^{PV}, tarifa)$ u trenutnom vremenskom koraku

Izlaz: skup uređaja koje treba uključiti u sledećem vremenskom koraku

```

1  Inicijalizacija:
   akcije  $\leftarrow \emptyset$ 
   proizvedena_struja_prelazi_prag  $\leftarrow S^{PV} > 0.1kWh$ 
   niska_tarifa  $\leftarrow tarifa == NISKA$ 
   for uređaj in uređaji do
2     if  $S_L^{netermalni}$  and not  $S_W^{netermalni}$  then
3       if  $S_p^{netermalni} == VISOK$  then
4         dodaj  $A_{uključiti}^{netermalni}$  u akcije
5       else if  $S_p^{netermalni} == SREDNJI$  then
6         if proizvedena_struja or tarifa  $\in \{NISKA, SREDNJA\}$  then
7           dodaj  $A_{uključiti}^{netermalni}$  u akcije
8         else if  $S_p^{netermalni} == NIJE VAŽNO$  and proizvedena_struja then
9           dodaj  $A_{uključiti}^{netermalni}$  u akcije
10      if  $S_p^{AC} == \{VISOK, SREDNJI\}$  then
11
```

```

12      if proizvedena_struja_prelazi_prag or niska_tarifa then
13          dodaj  $A_{uključujući}^{AC}$  u akcije
14      if  $S_{\Delta T_{razlika}} > 2.0$  and  $S_p^{AC} == HIGH$  then
15          dodaj  $A_{uključujući}^{AC}$  u akcije
16      end
17      return akcije

```

Algoritam 5 predstavlja sistem upravljanja radom uređaja zasnovan na znanju, koji je definisan od strane autora disertacije. Za razliku od prethodnog algoritma generisanog od strane LLM-a, ovaj pristup uvodi sofisticiraniju logiku prioritizacije i alokacije dostupne energije, oslanjajući se na detaljno razumevanje energetske i komfornih zahteva stanara.

Ulaz u algoritam 5 čini stanje sistema (S). Na izlazu algoritam vraća skup akcija uključivanja i isključivanja uređaja koje treba sprovesti u narednom vremenskom koraku. U algoritmu se prvo svi uređaji sortiraju prema značaju rada u datom vremenskom koraku, čime se obezbeđuje da se najpre razmatraju oni sa višim prioritetom. Uređaji se obrađuju sekvencijalno, a logika se razlikuje u zavisnosti od toga da li je u pitanju termalni (klima uređaj) ili netermalni uređaj (veš mašina i mašina za sudove):

Klima uređaj se uključuje ukoliko trenutna unutrašnja temperatura odstupa od željnog temperaturnog opsega, a značaj rada ovog uređaja je „Visok“. U suprotnom, klima uređaj se isključuje.

Netermalni uređaje koji rade ili nisu spremni za rad se preskaču (njihovo stanje rada se ne menja). Ponašanje za ostale netermalne uređaje zavisi od značaja njihovog rada i tarife:

- Uključuju se svi netermalni uređaji čiji je značaj „Visok“.
- Uređaji čiji je značaj „Srednji“ se uključuju ukoliko tarifa nije „Visoka“.

Svi ostali uređaji se isključuju.

Algoritam vodi računa da se uređaji uključuju tako da ne pređu trenutno dostupnu količinu solarne energije ako je moguće, a tek posle donosi odluke na osnovu značaja rada uređaja i tarife.

Algoritam 5. Sistem zasnovan na znanju za upravljanje uređajima, definisan od strane autora disertacije. Reprezentacija stanja S definisana je formulom (34), dok je skup akcija A definisan formulom (35):

- $(S_W^{uredaj}, S_L^{uredaj}, S_p^{uredaj})$ označavaju da li je uređaj već radio tokom dana, da li je uređaj spreman za rad i značaj rada uređaja
- E_t^{PV} označava proizvodnju solarnog panela u vremenskom koraku t .
- E_S označava ukupnu potrošenu energiju kada bi svi uređaji radili tokom vremenskog koraka t .
- *uređaji* predstavlja listu svih uređaja, sortiranu prema značaju njihovog rada u vremenskom koraku t .
- $E_{uredaj_maksimalna-potrošnja}$ označava maksimalnu potrošnju uređaja kada je uključen.
- *željena temperatura* je indikator koji označava da li se trenutna temperatura nalazi unutar željenog temperaturnog opsega.
- $A_{uključujući}^{uredaj}$ označava akciju uključivanja uređaja.
- $A_{isključujući}^{uredaj}$ označava akciju isključivanja uređaja.

Ulaz: Stanje $S = (S_{\Delta T_{razlika}}, S_p^{AC}, S_W^{CW}, S_L^{CW}, S_p^{CW}, S_W^{DW}, S_L^{DW}, S_p^{DW}, S^{PV}, tarifa)$ u trenutnom vremenskom koraku

Izlaz: kolekcija akcija koje treba sprovesti u sledećem vremenskom koraku

```

1  Inicijalizacija:
   akcije  $\leftarrow \emptyset$ 
    $E_S \leftarrow 0$ 
   uređaji  $\leftarrow$  uređaji sortirani na osnovu značaja rada
2  for uredaj in uređaji do
3      | if uredaj == AC then

```

```

4      if not željena_temperatura and  $S_p^{\text{uredaj}} == \text{VISOK}$  then
5          dodaj  $A_{\text{uključujući}}^{\text{uredaj}}$  u akcije
6      else
7          dodaj  $A_{\text{isključujući}}^{\text{uredaj}}$  u akcije
8  else
9      if  $S_W^{\text{uredaj}}$  or not  $S_L^{\text{uredaj}}$  then
10         continue
11     if  $E_t^{\text{PV}} > E_s$  and  $S_p^{\text{uredaj}} \neq \text{NIJE VAŽNO}$  then
12         dodaj  $A_{\text{uključujući}}^{\text{uredaj}}$  u akcije i
13         izvrši se  $E_s += E_{\text{maksimalnu-potrošnja}}^{\text{uredaja}}$ 
14     else
15         if  $S_p^{\text{uredaj}} == \text{VISOK}$  then
16             dodaj  $A_{\text{uključujući}}^{\text{uredaj}}$  u akcije i
17             izvrši se  $E_s += E_{\text{maksimalnu-potrošnja}}^{\text{uredaja}}$ 
18         else if  $S_p^{\text{uredaj}} == \text{SREDNJI}$  then
19             if tarifa  $\neq \text{VISOK}$  then
20                 dodaj  $A_{\text{uključujući}}^{\text{uredaj}}$  u akcije i
21                 izvrši se  $E_s += E_{\text{maksimalnu-potrošnja}}^{\text{uredaja}}$ 
22             else
23                 dodaj  $A_{\text{isključujući}}^{\text{uredaj}}$  u akcije
24     end
25 return akcije

```

Kod koji je definisan od strane domenskog eksperta složeniji je u poređenju sa kodom generisanog od strane LLM-a. Algoritam 4 vraća isključivo akcije za uključivanje uređaja, dok algoritam 5 obuhvata i akcije za njihovo isključivanje.

Za netermalne uređaje, algoritam 5 daje veći prioritet značaju rada. Ovaj algoritam ne dozvoljava rad netermalnih uređaja čiji je značaj „Nije važno“ i kada tarifa „Visoka“, dok algoritam 4 omogućava uključivanje čak i pri nižim prioritetima ukoliko postoji solarna proizvodnja ili povoljna tarifa.

Za termalne uređaje, algoritam 5 strogo zahteva uključivanje kada je temperatura van željenog opsega, dok algoritam 4 omogućava uključivanje na osnovu proizvedene solarne energije ili niske tarife, i kada je temperaturna razlika manja od praga.

Takođe, algoritam 5 dinamički prati kumulativnu potrošnju energije uključenih uređaja i usklađuje odluke sa raspoloživom solarnom proizvodnjom, dok algoritam 4 donosi odluke bez kontrole da li ukupna potrošnja premašuje trenutnu proizvodnju.

4.5.2 Few-shot strategija prompta

U okviru *few shot* strategije, u svakom koraku simulacije, LLM prima prompt sa opisom prethodnog stanja, preduzete akcije u tom stanju, neposredne nagrade koja je bila posledica te akcije i opis trenutnog stanja⁷. Modelu se daje instrukcija da na osnovu ovih

⁷ U prvom koraku simulacije, prompt ne sadrži primere prethodnih interakcija između okruženja i LLM-a. Kako simulacija odmiče, interakcije se sukcesivno memorišu i uključuju u naredne promptove, sve do n -tog koraka, nakon čega se memorija resetuje i započinje novo sekvencijalno prikupljanje do kraja dana. Ovakva strategija je usvojena kako bi se izbeglo prekomerno povećanje dužine prompta, koje može značajno usporiti

informacija generiše optimalnu akciju za trenutno stanje u predefinisanoj formatu. Ovim postupkom se imitira proces donošenja odluka u učenju potkrepljivanjem, pri čemu LLM koristi dostupne primere interakcija kako bi poboljšao svoje odluke u kontekstu optimizacije potrošnje energije.

Tabela 8 prikazuje standardizovani prompt sistemske poruke koji definiše ulogu modela i kontekst problema, te se dodaje na početak svakog korisničkog prompta. Tabela 9 prikazuje pojednostavljeni primer korisničkog prompta koji model prima u toku simulacije. Tabela 10 sadrži tekst koji se dodaje na kraj svakog korisničkog prompta. Ovaj tekst definiše željeni format odgovora i naglašava da je cilj maksimizacija nagrade, odnosno da vrednost nagrade bude pozitivna.

Dizajn prompta zasnovan je na preporukama iz literature. Radovi (Reif et al. 2022) i (Kirk et al. 2024) ističu da se efektivnost i konzistentnost odgovora LLM-a može poboljšati uključivanjem primera, graničnih oznaka (delimitera) i jasno označenih komponenti unutar prompta. U skladu sa preporukama iz radova (C. Yang et al. 2024) i (D. Zhou et al. 2023), prompt uključuje primere prethodnih koraka sa odgovarajućim rezultatima, ali je njihov broj ograničen na konačan broj uzastopno zabeleženih koraka zbog ograničenja kapaciteta tokena koje LLM može obraditi u jednom zahtevu. Sekvencijalni redosled primera u promptu, inspirisan primenom LLM-a u prognozi vremenskih serija (eng. *time series forecasting*) (Gruver et al. 2023) (Jin et al. 2024.), konceptualno odgovara mehanizmu ponavljanja iskustva (eng. *experience replay*) u dubokom učenju potkrepljivanjem. U dubokom učenju potkrepljivanjem, agent optimizuje politiku koristeći prethodno sačuvano iskustvo, dok LLM, iako ne ažurira svoju mrežu politike, koristi istorijske podatke o prethodnim sekvencijalnim odlukama za predviđanje budućih akcija. Ova paralela sa DRL strategijom omogućava modelu da koristi sekvencijalno i kontekstualno donošenje odluka za poboljšanje performansi.

Jedan od izazova u radu sa LLM-om bio je generisanje izlaza koji nisu mogli biti automatski parsirani, što je u skladu sa nalazima rada (Reif et al. 2022). Ovaj problem je ublažen primenom koncepta meta-promptova, kako je predloženo u radu (Y. Zhou et al. 2023), što je značajno smanjilo broj nepravilno formatiranih odgovora. U slučajevima kada LLM ipak generiše neispravan odgovor, on se tretira kao *Idle* akcija, odnosno, sistem ne preduzima nijednu radnju tokom tog vremenskog koraka. Dodatno, eksplicitno navođenje modela da svoj izlaz formatira u JSON strukturu na kraju svakog prompta doprinelo je većoj doslednosti i pouzdanosti parsiranja odgovora.

Tabela 8. Sistemska poruka za usmeravanje LLM-a

Prompt	Objašnjenje
You are a house energy management system. Your task is to optimize electricity consumption and user comfort in a simulated home, outputting an integer value from 0 to 6 for the chosen action.	Postavljanje tona prompta: objašnjenje cilja, izlaza i procesa simulacije.
The simulation progresses in 15-minute time steps. At each step, you will receive previous states, actions, and rewards for the action applied in that state. The reward reflects the quality of the action: a higher reward indicates a better action.	Opis vremenskog toka simulacije i mehanizma povratne informacije.

generisanje odgovora i degradirati tačnost modela, što je potvrđeno i u literaturi (Levy et al. 2024).

Restrictions: The Clothes Wash and Dishwasher can only operate once per day. Once started, they cannot be interrupted.	Postavljanje operativnih ograničenja za određene uređaje.
Action Dictionary: <pre> action_dict = { 0: 'AC: ON', 1: 'AC: OFF', 2: 'Clothes Wash: ON', 3: 'Clothes Wash: OFF', 4: 'Dishwasher: ON', 5: 'Dishwasher: OFF', 6: 'IDLE: DO NOTHING'} </pre>	Definisanje mogućih akcija koje sistem može preduzeti.
State Dictionary: The states are provided as a tuple of 10 values: <pre> state_dict = { 'Temp Difference': Difference between desired and actual indoor temperature (integer), 'Priority AC': ['High', 'Medium', 'Do not care'], 'Worked Today Clothes Wash': ['Worked Today', 'Didn't Work'], 'Loaded Clothes Wash': ['Loaded', 'Not Loaded'], 'Priority Clothes Wash': ['High', 'Medium', 'Do Not Care'], 'Worked Today Dishwasher': ['Worked Today', 'Didn't Work'], 'Loaded Dishwasher': ['Loaded', 'Not Loaded'], 'Priority Dishwasher': ['High', 'Medium', 'Do Not Care'], 'Production Pv': How much the PV produced, 'Tariff': ['Off Shoulder', 'Shoulder', 'Peak'], 'Time': Current time} </pre>	Objašnjenje reprezentacije stanja korišćene u simulaciji.
Priority Levels: - 'High': The device must work, - 'Medium': The device should work, - 'Do Not Care': It does not matter if the device works.	Definisanje nivoa prioriteta za različite akcije.
The simulation begins at 00:00 (midnight) and runs until 00:00 (midnight) the following day.	Definisanje vremenskog okvira simulacije i glavnog zadatka HEMS-a.
Your task is to determine the best action for the given state. To determine the action, do the following steps: - Look at the previous states, actions you took, and the rewards that resulted from taking a certain action. - Always strive to get a positive reward for each time step. Do not decide on the next action before considering the restrictions. - Check the defined priority levels. If they are high and still do not work, turn the device on - Check the PV production. If the number is greater than 50, consider turning on a device that didn't work or has a high priority - Check the tariff price. Consider turning off devices when the tariff is peak	Navođenje detaljnih koraka za određivanje optimalne akcije u datom stanju.

Use the following format to evaluate the next possible action: The Previous states, actions, and the rewards: " state, action, reward for each time step here " Current state: " Current state here "	Formatiranje odgovora.
---	------------------------

Tabela 9. Uprošćen primer korisničkog prompta

Prompt	Odgovor LLM-a	Objašnjenje
The previous states, actions, and the rewards: " <other previous states, actions and rewards> Temp Difference: 0 Priority Ac: Do Not Care Worked_Today Clothes Wash: Didn't Work Loaded Clothes Wash: Loaded Priority Clothes Wash: Do Not Care Worked Today Dishwasher: Didn't Work Loaded Dishwasher: Loaded Priority Dishwasher: Medium Production Pv: -204 Tariff: Shoulder Time: 13:15 For the action: Idle The reward was: 0.4499999999999996 " The Current State: " Temp Difference: 0 Priority Ac: Do Not Care Worked Today Clothes Wash: Didnt Work Loaded Clothes Wash: Loaded Priority Clothes Wash: Do Not Care Worked Today Dishwasher: Didnt Work Loaded Dishwasher: Loaded Priority Dishwasher: Medium Production Pv: -184 Tariff: Shoulder Time: 13:30 "	{"actions": [4]}.	Mašina za sudove nije radila tokom dana i trenutno je spremna za upotrebu (napunjena sudovima). Nivo prioriteta dodeljen mašini za sudove je srednji, a tarifa električne energije se nalazi u srednjem cenovnom periodu. Pored toga, u ovom trenutku je zabeležena proizvodnja električne energije putem solarnog panela. Na osnovu ovih uslova, LLM je predložio akciju 4 – uključivanje mašine za sudove.

The previous states, actions, and the rewards: "" <other previous states, actions and rewards> Temp Difference: 2 Priority Ac: High Worked_Today Clothes Wash: Didnt Work Loaded Clothes Wash: Loaded Priority Clothes Wash: Do Not Care Worked_Today Dishwasher: Worked Today Loaded Dishwasher: Not Loaded Priority Dishwasher: Do Not Care Production Pv: 0 Tariff: Shoulder Time: 18:45 For the action: AC: on The reward was: 93.68653112989435 "" The current state: "" Temp Difference: 0 Priority Ac: High Worked_Today Clothes Wash: Didnt Work Loaded Clothes Wash: Loaded Priority Clothes Wash: Medium Worked_Today Dishwasher: Worked Today Loaded Dishwasher: Not Loaded Priority Dishwasher: Do Not Care Production Pv: 0 Tariff: Shoulder Time: 19:00 ""	{"actions": [2]}.	Mašina za veš nije bila aktivna tokom dana i trenutno je spremna za upotrebu (napunjena vešom). Njen nivo prioriteta je srednji, a tarifa električne energije se nalazi u srednjem cenovnom periodu. Na osnovu ovih uslova, LLM je predložio akciju 2 – uključivanje mašine za veš.
---	--------------------------	--

Tabela 10. Dodatni tekst na kraju korisničkog prompta

Prompt	Objašnjenje
Respond only in the form of '{ "actions": [] }'. Actions are int values from 0 to 6. The goal is for the reward to be a positive value.	Definisanje željenog formata izlaza.

5 Evaluacija metodologije

Ovo poglavlje opisuje eksperimentalnu postavku za evaluaciju razvijenih HEMS pristupa i prikazuje dobijene rezultate evaluacije. Potpoglavljje 5.1 predstavlja studiju slučaja, definišući parametre simulacije pametne kuće koja se koristi za razvoj i evaluaciju HEMS sistema.

5.1 Studija slučaja: eksperimentalna postavka za simulaciju pametne kuće

U okviru disertacije dizajnirano je adaptivno simulaciono okruženje koje oponaša potrošnju električne energije u pametnoj kući (potpoglavljje 4.2). U ovom potpoglavljju, to okruženje je korišćeno za simulaciju studije slučaja – pametne kuće locirane u Nemačkoj, koja je opremljena solarnim panelom. Od uređaja kontrolisanih od strane HEMS-a, prisutni su klima uređaj, mašina za sudove i mašina za veš (tabela 11). Za svaki od ovih uređaja definisani su zahtevi u pogledu željenog vremena rada i temperature unutrašnjeg prostora. Ova postavka simulacije korišćena je za obučavanje i evaluaciju različitih HEMS pristupa.

Tabela 11. Tipovi uređaja koji su podržani u simulacionim okruženjima.

Uređaj/Tip uređaja	Promenljivo vreme rada	Rad zavisian od eksternih vremenskih podataka	Potrošač energije	Proizvodač energije
Klima uređaj	✓	✓	✓	✗
Mašina za sudove	✓	✗	✓	✗
Mašine za veš	✓	✗	✓	✗
Solarni panel	✗	✓	✗	✓

Potpoglavljje 5.1.1 prikazuje karakteristike uređaja u pametnoj kući i akcije kojima ih HEMS može kontrolisati. Takođe je istaknuta razlika u simulaciji ovih uređaja u zavisnosti od korišćenog simulacionog okruženja, odnosno od toga da li se njihova potrošnja modeluje kao konstantna ili varijabilna. Potpoglavljje 5.1.2 prikazuje eksterne podatke relevantne za simulaciju pametne kuće na određenoj lokaciji, uključujući istorijske podatke o vremenskim uslovima i dostupnim tarifama. Potpoglavljje 5.1.3 specificira zahteve stanara u vezi sa radom uređaja koje kontroliše HEMS. Potpoglavljje 5.1.4 sumira definiciju stanja i akcija dostupnih HEMS-u. Konačno, potpoglavljje 5.1.5 definiše eksperimentalnu postavku.

5.1.1 Modelovanje uređaja u pametnoj kući

Svaki od uređaja definisanih u narednim potpoglavljima modelovan je korišćenjem *Typhoon HIL* (Typhoon HIL 2008) i *Load Flow* (Vojnović et al. 2022) (Vojnović et al. 2023b) (Vojnović et al. 2023a) okruženja. Razlika između ovih okruženja ogleda se u načinu modelovanja potrošnje električne energije tokom rada uređaja. U okviru *Typhoon HIL* simulacionog okruženja, potrošnja uređaja se modeluje kao konstantna – vrednost potrošnje u svakom vremenskom koraku predstavlja prosečnu vrednost potrošnje tokom

celokupnog vremena rada uređaja. Nasuprot tome, *Load Flow* okruženje omogućava modelovanje promenljive potrošnje uređaja tokom rada, čime se postiže realističnija simulacija njihovog rada. Modelovanje uređaja pomoću *Load Flow* simulacionog okruženja je opisano u potpoglavlju 2.1.1.

Svaki od modelovanih uređaja ima unapred definisan skup akcija kojima HEMS može upravljati njihovim radom. Osnovna akcija dostupna za sve HEMS-kontrolisane uređaja jeste uključivanje ili isključivanje na početku vremenskog koraka, što određuje da li će uređaj raditi tokom tog vremenskog intervala. Na kraju svakog vremenskog koraka, HEMS može očitati izlazne podatke uređaja, koji su relevantni za razumevanje trenutnog stanja sistema i donošenje narednih odluka.

5.1.1.1 Klima uređaj

Klima uređaj (eng. *Air Condition*, AC) je modelovan kao termalni uređaj sa promenljivim vremenom rada, i predstavlja potrošač energije.

Unutrašnja temperatura kuće zavisi ne samo od rada klima uređaja, već i od termodinamičkih karakteristika kuće i spoljašnjih vremenskih uslova. Termodinamika kuće obuhvata razmenu toplote sa spoljašnjim okruženjem kroz zidove i prozore, uključujući gubitke i dobitke toplote.

Na početku svakog vremenskog koraka, HEMS može da izvrši sledeće akcije:

- Uključi AC – Klima uređaj započinje cikluse hlađenja ili grejanja kako bi održao unutrašnju temperaturu unutar željenog intervala od 20°C do 25°C.
- Isključi AC – Klima uređaj prestaje sa radom, čime prestaje i regulacija unutrašnje temperature.

Potrošnja energija može se optimizovati regulacijom unutrašnje temperature u skladu sa prisustvom stanara u domaćinstvu. Idealni scenario podrazumeva isključenje klima uređaja pre odlaska stanara iz kuće, uz očuvanje temperature unutar željenog opsega dok su prisutni. Nakon njihovog odlaska, temperatura bi se postepeno prilagođavala spoljašnjim vremenskim uslovima. Klima uređaj bi se ponovo aktivirao pre povratka stanara, kako bi osigurao postizanje željene unutrašnje temperature pre njihovog dolaska. Ovakav pristup sprečava naglu i visoku potrošnju električne energije potrebne za brzo vraćanje temperature u željeni opseg, što bi se moglo desiti u slučaju većih temperaturnih odstupanja u trenutku povratka stanara.

Na kraju svakog vremenskog koraka, klima uređaj vraća informaciju o potrošnji električne energije P_{AC} . U zavisnosti od korišćenog okruženja, izlazne vrednosti potrošnje mogu se razlikovati:

- *Typhoon HIL* simulaciono okruženje: vraća konstantnu potrošnju od 1.8 kW (Pedrasa, Spooner, and MacGill 2010).
- *Load Flow* simulaciono okruženje: vraća promenljivu potrošnju izraženu u kW, kao što je opisano u potpoglavlju 2.1.1.

Zbog različite snage električne energije koju koristi klima uređaj u zavisnosti od simulacionog okruženja, gde uređaj u *Typhoon HIL* ima slabiju snagu u poređenju sa uređajem u *Load Flow*, temperatura sporije raste tokom zime, odnosno sporije opada tokom leta.

5.1.1.2 Mašina za sudove

Mašina za sudove (eng. *dishwasher*) je uređaj sa promenljivim vremenom rada i potrošač energije sa neprekidnim radnim ciklusom, čija potrošnja energije u realnim uslovima varira tokom različitih faza rada. Ukoliko dođe do prekida rada, HEMS mora pokrenuti pranje ispočetka, pri čemu sva do tada utrošena energija predstavlja nepovratni gubitak. Nakon ovakvog prekida, uređaj ostaje u stanju „spreman za rad“. Zbog ovih karakteristika, neophodno je pažljivo planirati vreme rada mašine za sudove, kako bi se istovremeno očuvao komfor stanara u pogledu željenog vremena završetka pranja i optimizovali troškovi električne energije, kroz usklađivanje rada uređaja sa proizvodnjom solarnog panela i važećim tarifnim režimima.

Na početku svakog vremenskog koraka t , HEMS može da izvrši jednu od sledećih akcija u vezi sa mašinom za sudove:

- Pokrene uređaj, ukoliko je spreman za rad. Uređaj se smatra spremnim za rad ako je napunjen sudovima. Ukoliko uređaj nije spreman, njegovo pokretanje nije moguće. U okviru simulacije razvijene u ovom istraživanju, uređaj je spreman za rad samo jednom dnevno – na početku dana. Po završetku pranja, uređaj se više ne smatra spremnim za rad do narednog dana.
- Zaustavi uređaj, čime se prekida njegov rad, čime se trenutni radni ciklus prekida. U tom slučaju, pranje sudova moraće da se započne ispočetka, a do tada utrošena energija se gubi.

Na kraju svakog vremenskog koraka t , mašina za sudove vraća sledeće izlazne vrednosti:

- $A_{dishwasher}$ – Indikator koji označava da li je uređaj spreman za rad.
- $P_{dishwasher}$ – Potrošnja električne energije uređaja tokom vremenskog koraka t izražena u kW i varira u zavisnosti od korišćenog okruženja:
 - *Typhoon HIL* simulaciono okruženje: vraća konstantnu potrošnju od 1 kW , gde je ukupan vreme rada 1h i 30min, odnosno 6 ciklusa od 15 minuta (Pedrasa, Spooner, and MacGill 2010).
 - *Load Flow* simulaciono okruženje: vraća promenljivu potrošnju u kW , gde je ukupno vreme rada 2h, odnosno 8 ciklusa od 15 minuta.

5.1.1.3 Mašina za veš

Kombinovana mašina za pranje i sušenje veša (eng. *washing machine*) predstavlja uređaj promenljivog vremena rada koji je potrošač energije. Na početku svakog vremenskog koraka t , HEMS može izvršiti akcije:

- Uključi uređaj.
- Zaustavi uređaj.

Na kraju svakog vremenskog koraka t , mašina za veš vraća sledeće izlazne vrednosti, analogno mašini za sudove:

- $A_{washingmachine}$ – indikator spremnosti uređaja za rad. U simulacijama sprovedenim u okviru ovog istraživanja, uređaj se smatra spremnim za rad isključivo na početku dana. Ukoliko uređaj nije spreman, njegovo pokretanje nije moguće. Po završetku pranja, uređaj se više ne smatra spremnim za rad do narednog dana.

- $P_{\text{washingmachine}}$ – potrošnja električne energije mašine za veš u vremenskom koraku t , izražena u kilovatima (kW). Tip potrošnje zavisi od korišćenog simulacionog okruženja:
 - *Typhoon HIL* simulaciono okruženje: modeluje potrošnju kao konstantnu, sa vrednošću od $0,5 kW$ i ukupnim vremenom rada od 2 sata (Pedrasa, Spooner, and MacGill 2010).
 - *Load Flow* simulaciono okruženje: omogućava modelovanje promenljive potrošnje tokom ciklusa rada. U ovom slučaju, ukupno vreme rada iznosi 4 sata i 30 minuta, odnosno 18 vremenskih koraka od po 15 minuta.

5.1.1.4 Solarni panel

Solarni panel je proizvođač električne energije čiji rad zavisi od eksternih vremenskih uslova. Proizvodnja električne energije solarnog panela izračunava se na osnovu stvarnih meteoroloških podataka o spoljašnjoj temperaturi i solarnoj iradijaciji, koji su detaljno opisani u potpoglavlju 5.1.2. Proizvedena energija može direktno napajati potrošače u domaćinstvu, pri čemu se višak energije ne može prodavati električnoj mreži. Na kraju svakog vremenskog koraka t , solarni panel vraća vrednost trenutne snage proizvedene električne energije P_{PV} , izraženu u kilovatima (kW), nezavisno od korišćenog simulacionog okruženja.

5.1.2 Vremenski uslovi i tarifni režimi

Definisan tarifni režim (tabela 12) oslikava domaćinstvo iz Nemačke⁸ i predstavljen je u €/kWh.

Tabela 12. Pregled tarifnih režima.

	Sati	Cena (€/kWh)
Visoka cena struje	5h – 13h	0,4
Srednja cena struje	13h – 00h	0,3
Niska cena struje	00h – 5h	0,2

Podaci o vremenskim uslovima, iradijaciji i spoljašnjoj temperaturi preuzeti su sa sajta Solcast⁹. Podaci su prikupljeni za Berlin na 15 minutnom vremenskom intervalu tokom dva različita perioda – od 1. do 23. januara i juna 2016.

5.1.3 Definisanje zahteva stanara pametne kuće

Ograničenje koje postavljaju stanari odnosi se na željeni opseg unutrašnje temperature, definisan u intervalu od $20^{\circ}C$ do $25^{\circ}C$. Pored toga, komfor stanara je modelovan kroz njihove zahteve u pogledu vremenskog rasporeda rada pojedinačnih uređaja. Vrednosti značaja režima rada koje stanari koriste za definisanje prioriteta rada uređaja (prikazane u tabeli 13) određuju se u skladu sa tarifnim režimima za električnu energiju prikazanim u tabeli 12, na osnovu pristupa (Pedrasa, Spooner, and MacGill 2010).

⁸ <https://www.agora-energiewende.de/>

⁹ <https://solcast.com/>

Radi objedinjavanja sa troškovima električne energije, komfor stanara kvantifikovan je u monetarnim jedinicama (€/kWh). Monetarne vrednosti dodeljene značaju rada uređaja (prikazane u tabeli 14) formulisane su tako da odražavaju spremnost stanara da koriste određene uređaje u specifičnim tarifnim periodima:

- Visoka vrednost dodeljuje se uslugama koje stanari očekuju i u periodima visoke tarife, te je viša od maksimalne cene električne energije.
- Srednja vrednost nalazi se između tarife srednjeg i visokog tarifnog perioda, što sugeriše da se upotreba tih uređaja može vremenski prilagođavati.

Tabela 13 prikazuje dodeljene prioritete za svaki uređaj tokom različitih perioda dana. Navedene vrednosti predstavljaju ulazne parametre simulacije, koje stanari definišu radi iskazivanja svojih preferencija u pogledu željenog vremena rada uređaja.

Tabela 13. Značaj rada uređaja u različitim periodima dana prema korisničkim zahtevima.

Uređaj	Režim rada	Značaj režima rada
Klima uređaj	17h – 8h (naredni dan)	Visok
	8h – 17h	Nije važno
Mašina za sudove	9h – 14h	Srednji
	14h – 17h	Visok
	Ostatak dana	Nije važno
Mašina za veš	19h – 21h	Srednji
	21h – 24h	Visok
	Ostatak dana	Nije važno

Tabela 14. Novčane vrednosti dodeljene radu uređaja na osnovu korisničkih prioriteta.

Značaj rada	Vrednost (€/kWh)	Opis
Visok	0,45	Kritična usluga koju je poželjno pružiti čak i tokom perioda visoke tarife. Vrednost je postavljena iznad maksimalne tarife kako bi se obezbedio prioritet rada uređaja.
Srednji	0,35	Usluga čiji se rad može ograničiti tokom perioda više tarife. Vrednost je između tarifa za srednji i visok tarifni period, odražavajući umeren prioritet.
Nije važno	Uređaji koji nisu termalni: 0 Termalni uređaji: 0,1	Stanari su ravnodušni prema radu uređaja u datom vremenskom periodu. Kod termalnih uređaja dodeljuje se mala nenulta vrednost, kako bi se penalizovala značajna odstupanja od željenog temperaturnog opsega i sprečilo prekomerno hlađenje ili zagrevanje unutrašnjeg prostora.

U primerima 5.1 i 5.2 prikazana je logika definisanja vremenskih opsega značaja rada za mašinu za veš i mašinu za sudove, u skladu sa zahtevima stanara. Primer 5.3, na primeru klima uređaja, ilustruje kako se, na osnovu zadatih prioriteta i njima pridruženih monetarnih vrednosti, izračunava vrednost komfora u zavisnosti od perioda rada uređaja.

Primer 5.1. Postavka značaja komfora na primeru veš mašine.

Pretpostavimo da stanari žele da veš bude opran do 21 časa. Značaj rada veš mašine varira tokom dana, u zavisnosti od preostalog vremena za završetak ciklusa:

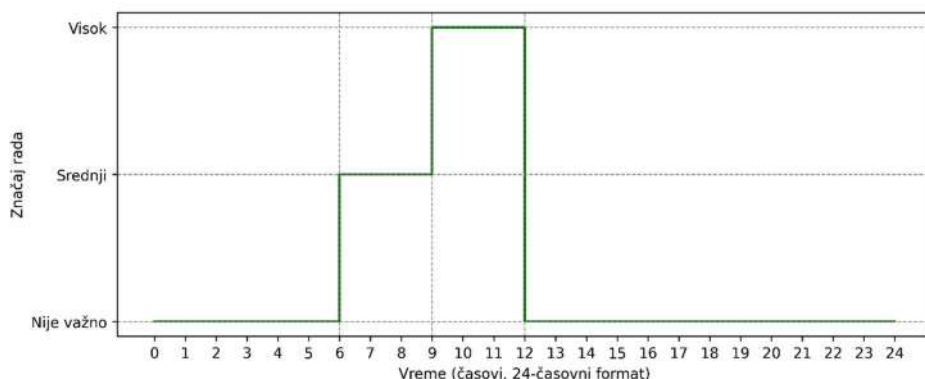
- Na početku dana, značaj je postavljen na „Nije važno”, jer nije bitno kada će se veš prati, sve dok bude gotov do 21 časa.
- Kada se približava 21 čas, značaj prelazi u „Srednji“, signalizirajući da bi uskoro trebalo pokrenuti uređaj.

- U 21 čas, kada ostane tačno onoliko vremena koliko je potrebno za završetak ciklusa pre ponoći, značaj se postavlja na „Visok”, jer je to poslednja mogućnost za ispunjenje cilja.

Primer 5.2. Postavka značaja komfora na primeru mašine za sudove.

Pretpostavlja se da stanari žele da sudovi budu oprani najkasnije do 12 časova (slika 19). Značaj rada mašine za sudove varira tokom dana, u skladu sa ovim zahtevom:

- Na početku dana, značaj je postavljen na „Nije važno“, jer vreme pokretanja nije kritično sve dok se ciklus može završiti do 12 časova.
- Od 6 do 8 časova, značaj prelazi u „Srednji“, signalizirajući da se približava trenutak kada bi uređaj trebalo aktivirati, ali još postoji fleksibilnost u planiranju vremena pokretanja.
- Od 8 časova, značaj postiže nivo „Visok“, jer je preostalo tačno onoliko vremena koliko je potrebno da se ciklus pranja završi do 12 časova. Uređaj bi u ovom trenutku trebalo pokrenuti, čak i ako je cena električne energije visoka.
- Nakon 12 časova, značaj se vraća na „Nije važno“, pod pretpostavkom da je zahtev korisnika ispunjen. Ukoliko uređaj nije pokrenut na vreme, pretpostavlja se da korisnicima više nije bitno kada će biti aktiviran.



Slika 19. Primer dinamičkog određivanja značaja rada mašine za sudove.

Primer 5.3. Postavka računanja komfora na primeru klima uređaja.

U ovom primeru analiziran je klima uređaj kao ilustracija računanja komfora prema izrazu (27). Radi pojednostavljenja, pretpostavlja se da je vrednost penalizacije $Kazna_t^i = 1$.

Tabela 15 prikazuje vrednosti komfora za klima uređaj za svaki 15-minutni vremenski korak. HEMS upravlja unutrašnjom temperaturom kontrolisanjem rada klima uređaja, s ciljem održavanja temperature unutar zadatog temperaturnog opsega. Uređaj se može isključiti, što omogućava prirodni porast ili pad unutrašnje temperature, u zavisnosti od spoljašnjih vremenskih uslova.

Stanari definišu visok značaj rada klima uređaja u periodima kada su prisutni u kući – u ovom primeru, to je interval od 17:00 do 8:00 narednog dana. Tokom odsustva, značaj rada uređaja se postavlja na „Nije važno“. Ipak, kako bi se izbegao prekomerni pad ili rast temperature dok su stanari odsutni, termalnim uređajima se i u tom periodu dodeljuje

mala nenulta novčana vrednost značaja, koja u ovom primeru iznosi 0,1 €/kWh, u skladu sa vrednostima prikazanim u tabeli 14. Radi jednostavnosti prikaza kazna iz formule (27) se ne uzima u razmatranje.

Tabela 15. Primer izračunavanja novčanog značaja rada klima uređaja tokom dana. Vrednost B_t^{AC} predstavlja novčani značaj rada klima uređaja u vremenskom koraku t , u skladu sa opisom iz tabele 14. Vrednost E_t^{AC} označava potrošnju električne energije klima uređaja u istom vremenskom koraku. Podaci o trenutnoj potrošnji preuzeti su iz simulacionog okruženja *Typhoon HIL*, opisanog u potpoglavlju 5.1.1.1. Komfort koji obezbeđuje klima uređaj modelovan je pomoću pojednostavljene jednačine (27), pri čemu se razmatra isključivo rad klima uređaja, bez penalizacije, uz pretpostavku da je potrošnja u svakom vremenskom koraku različita od nule ($E_t^{AC} \neq 0$).

Vreme	B_t^{AC}	E_t^{AC}	Komfor AC
7:45	0,45 (Visok)	1.8 kW * 15min (0,25h) = 0,45 kWh	0,45 * 0,45 = 0,2025
8:00	0,1 (Nije važno)	1.8 kW * 15min (0,25h) = 0,45 kWh	0,45 * 0,1 = 0,045
8:15	0,1 (Nije važno)	1.8 kW * 15min (0,25h) = 0,45 kWh	0,45 * 0,1 = 0,045
...			
16:45	0,1 (Nije važno)	1.8 kW * 15min (0,25h) = 0,45 kWh	0,45 * 0,1 = 0,045
17:00	0,45 (Visok)	1.8 kW * 15min (0,25h) = 0,45 kWh	0,4 * 0,45 = 0,2025

5.1.4 Reprezentacija stanja i skup mogućih akcija

U ovom potpoglavlju opisane su ključne komponente simulacionog okruženja koje definišu simulaciju u svakom vremenskom koraku – reprezentacija posmatranog stanja (potpoglavlje 5.1.4.1) i moguće akcije koje HEMS može preduzeti (potpoglavlje 5.1.4.2).

5.1.4.1 Reprezentacije stanja

Reprezentacija stanja zavisi od prirode posmatranog okruženja. Na primer, u video-igramama stanje može biti predstavljeno jednim frejmom igre (eng. *game frame*) (Mnih et al. 2013). U kontekstu HEMS-a prikazanog u ovoj disertaciji, stanje obuhvata: izlazne vrednosti uređaja definisanih u potpoglavlju 5.1.1, vremenske uslove i tarifni režim iz potpoglavlja 5.1.2, kao i zahteve stanara definisane u potpoglavlju 5.1.3. Stanje je modelovano kao vektor diskretnih, celobrojnih i realnih vrednosti:

$$S = (S_{\Delta T_{\text{razlika}}}, S_p^{AC}, S_W^{CW}, S_L^{CW}, S_p^{CW}, S_W^{DW}, S_L^{DW}, S_p^{DW}, S^{PV}, \text{tarifa}), \quad (34)$$

gde su komponente vektora definisane na sledeći način:

- $S_{\Delta T_{\text{razlika}}}$ predstavlja razliku između unutrašnje temperature na početku trenutnog vremenskog koraka (za koji se donosi odluka) i ciljanog temperaturnog opsega definisanog od strane korisnika. Vrednosti su celobrojne pozitivne vrednosti.
- S_p^i označava nivo značaja rada uređaja i na početku trenutnog vremenskog koraka („Visok“, „Srednji“ ili „Nije važno“, prema tabeli 13) za sve uređaje kojima HEMS upravlja: klima uređaj (AC), mašina za pranje veša (CW) i mašina za pranje sudova (DW). Radi stabilizacije procesa obučavanja modela, kategorije su predstavljene kao celobrojne vrednosti: 2 za „Visok“, 1 za „Srednji“ i 0 za „Nije važno“. Pored ovoga klima uređaj u kategoriji „Nije važno“ ima vrednost 0,1.

- Za netremalne uređaje kojima HEMS upravlja (CW i DW):
 - S_W označava da li je uređaj bio aktivan u prethodnom vremenskom koraku.
 - S_L označava da li je uređaj spreman za rad na početku trenutnog vremenskog koraka.
- S_{PV} označava količinu energije koju je solarni panel proizveo u tekućem vremenskom koraku. Ova vrednost procenjenjuje se na osnovu solarne iradijacije izmerene na početku tog koraka. Vrednosti su prikazane kao negativni celi brojevi.
- *tarifa* označava režim cene električne energije na početku trenutnog vremenskog koraka, koji može poprimiti jednu od tri diskretne vrednosti prikazane u tabeli 12. Radi stabilizacije procesa obučavanja modela, kategorije su predstavljene kao celobrojne vrednosti: 2 za „Visoka cena struje“, 1 za „Srednja cena struje“ i 0 za „Niska cena struje“.

5.1.4.2 Skup akcija

Definisani skup akcija za uređaje kojima HEMS može upravljati je binaran:

$$A = \{A_{uključi}^i, A_{isključi}^i, A_{idle} | i \in \{1, \dots, N\}\}, \quad (35)$$

gde je: A skup mogućih akcija, $A_{uključi}^i$ akcija uključivanja uređaja i , $A_{isključi}^i$ akcija isključivanja uređaja i , a N ukupan broj uređaja kojima HEMS upravlja. Dodatna akcija kada agent ne preduzima nikakvu akciju A_{idle} . Ukupno ima 7 akcija.

Zbog ograničenja modela, pristupi zasnovani na učenju potkrepljivanjem mogu odaberati samo jednu optimalnu akciju u vremenskom koraku. Nasuprot tome, LLM modeli omogućavaju odabir više akcija.

5.1.5 Eksperimentalna postavka

Trening skup za modele učenja potkrepljivanjem obuhvata prve tri nedelje u mesecu, odnosno prvih 22 dana. Tokom obuke, za svaki trening korak nasumično se bira jedan od dana iz trening skupa, čime se izbegava moguća zavisnost između uzastopnih dana i podstiče sposobnost generalizacije modela. Za razliku od modela učenja potkrepljivanjem, LLM modeli ne zahtevaju proces obučavanja, već se primenjuju direktno u inferenciji.

Zbog ograničenja dostupnog budžeta ovom istraživanju, evaluacija LLM modela sprovedena je samo na jednom danu – 23. dan u mesecu. Radi objektivnog poređenja, svi modeli su evaluirani na istom danu, uključujući:

- Modele učenja potkrepljivanjem (DDQN i PPO),
- Sistem baziran na znanju (pravila definisana od strane autora rada) (KBS stručnjak),
- Modeli bazirani na velikim jezičkim modelima:
 - KBS definisan od strane LLM bez primera (KBS LLM),
 - LLM sa nekoliko primera simulirajući ponašanje RL-a (RL LLM).

Prikazana je evaluacija modela primenom metrika za komfor (27), potrošnju (32) i ukupnu nagradu (33). Detaljno će se analizirati i uporediti odluke različitih modela, gde će se istaći najbolji postignuti rezultati.

Dodatno, modeli DDQN, PPO i KBS definisan od strane ljudskog eksperta i KBS definisan od strane LLM bez primera i LLMM sa nekoliko primera simulirajući ponašnje RL-a su evaluirani na 7 dana, od 23. do 31. dana u januaru i junu, da bi se ispitala konzistentnost dobijenih rezultata.

Radi provere robusnosti modela, opisana evaluacija je sprovedena u dva eksperimenta, za dva različita meseca koji predstavljaju različita godišnja doba (zima i leto). Ovaj pristup osigurava da su modeli izloženi različitim uslovima potrošnje električne energije i vremenskim uslovima, čime se ispituje njihova otpornost na sezonske promene.

U ovom potpoglavlju detaljno se analiziraju rezultati modela dubokog učenja. U potpoglavlju 6.1 analiziraju se rezultati DDQN i PPO modela, gde su rezultati predstavljeni u kontekstu simulacionih okruženja *Typhoon HIL* i *Load Flow*. U potpoglavlju 6.2 prikazana je analiza rezultata velikog jezičkog modela.

Prvo su analizirane vrednosti funkcija nagrade i odluke modela na jednom danu koji je sledio nakon perioda treniranja. Zatim su, za potrebe ispitivanja robusnosti modela, analizirane vrednosti funkcije nagrade u periodu od sedam dana koji slede nakon perioda treniranja. Ista analiza sprovedena je u zimskom i letnjem periodu, kako bi se sagledale sezonske varijacije u rezultatima.

6 Rezultati i diskusija

U ovom poglavlju prikazani su i diskutovani rezultati modela za orkestraciju pametnih uređaja sa ciljem optimizacije udobnosti stanara i potrošnje električne energije. U potpoglavlju 6.1 analiziraju se rezultati modela učenja potkrepljivanjem uz detaljnu analizu procesa obučavanja i odluka obučanih modela na jednom testnom danu koji sledi neposredno nakon perioda treninga. Potpoglavlje 6.2 obuhvata istu analizu u kontekstu primene velikih jezičkih modela. U potpoglavlju 6.3 rezultati svih modela su upoređeni sa performansama sistema zasnovanog na znanju definisanog od strane ljudskog eksperta. U potpoglavlju 6.4 ispituje se robusnost modela pri donošenju odluka u periodu od sedam uzastopnih dana. Na kraju, potpoglavlje 6.5 sumira i diskutuje sve rezultate.

6.1 Analiza rezultata modela učenja potkrepljivanjem

U ovom potpoglavlju prikazani su rezultati ostvareni pomoću modela dubokog učenja potkrepljivanjem. Potpoglavlje 6.1.1 analizira proces konvergencije nagrade – prikazuje se promena vrednosti srednje dnevne nagrade tokom obuke radi utvrđivanja da li su modeli konvergirali. U potpoglavlju 6.1.2 prikazane su vrednosti funkcije nagrade koji trenirani modeli ostvaruju na testnom danu koji sledi neposredno nakon perioda obuke. Za svako simulaciono okruženje prikazana je posebna tabela, pri čemu su u svakoj tabeli objedinjeni rezultati za zimski i letnji period. Potpoglavlje 6.1.3 analizira odluke modela na istim testnim danima. Ovde su zasebno analizirane odluke modela vezane za kontrolisanje unutrašnje temperature i upravljanje radom mašine za pranje sudova i mašine za veš.

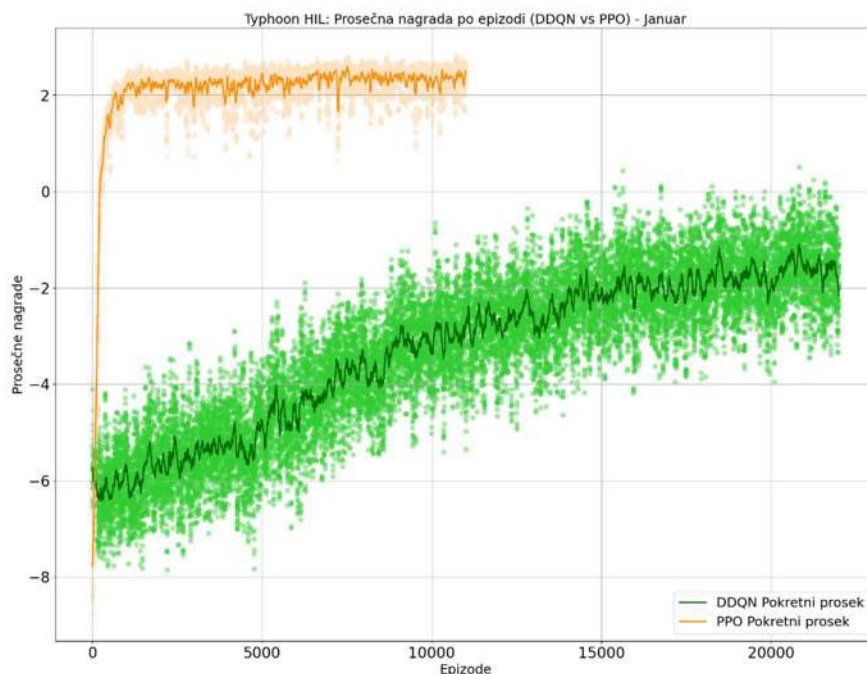
6.1.1 Analiza procesa konvergencije nagrade

Prikaz rezultata konvergencije modela ističe brzinu obučavanja modela, gde se od istreniranog modela očekuje stabilan rast vrednosti nagrade kroz iteracije, uz minimalne oscilacije. DDQN i PPO modeli su trenirani u dva različita meseca iz 2016. godine (potpoglavlje 5.1.2) – zimskom (od 01. do 22. januara) i letnjem (od 01. do 22. juna). Krive obučavanja DDQN i PPO modela u simulacionim okruženjima *Typhoon HIL* i *Load Flow* za ove mesece su prikazane na slikama 20 i 21.

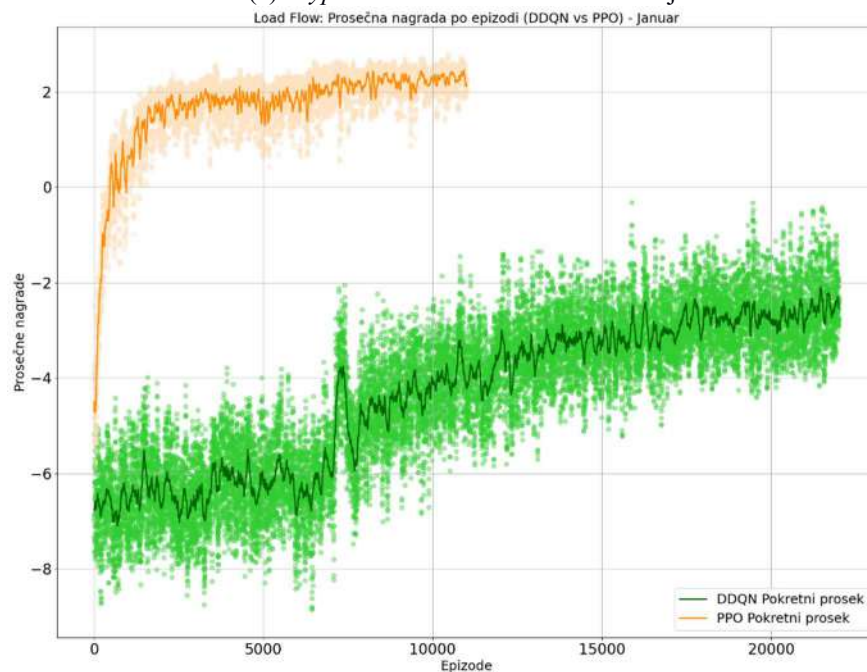
Modeli učenja potkrepljivanjem zahtevaju veliki broj iteracija kako bi dostigli stabilno ponašanje (Cobbe et al. 2019). Bez obzira na scenario, PPO modeli su se obučili nakon 11.000 epizoda u zimskom i letnjem periodu. Za razliku od njih, DDQN modeli su zahtevali veći broj epizoda za uspešno obučavanje (oko 22.000 epizoda u zimskom i 40.000 epizoda u letnjem periodu).

Konačne prosečne dnevne nagrade DDQN modela bile su niže i pokazivale su veće oscilacije u poređenju sa nagradama PPO modela. U zimskom periodu, u oba okruženja, PPO je ostvario prosečnu dnevnu nagradu od oko 2 €, dok je DDQN postigao oko -2 €. U letnjem periodu, u *Typhoon HIL* simulacionom okruženju, prosečna dnevna nagrada PPO modela iznosila je oko 3 €, dok je DDQN postigao približno 0 €. U letnjem periodu u *Load Flow* simulacionom okruženju, prosečna dnevna nagrada PPO modela iznosila je oko 2 €, dok je DDQN postigao prosečnu dnevnu nagradu od oko 0 €.

Ovde prikazani rezultati su u skladu sa rezultatima prikazanim u radovima (Kozlica et al. 2023) (Sabbir M 2024), koji su poredili performanse DDQN i PPO modela u rešavanju različitih problema. Rad (Kozlica et al. 2023) je utvrdio da DDQN model sporije konvergira u poređenju sa PPO modelom, kao i da PPO postiže bolje rezultate nakon treniranja.

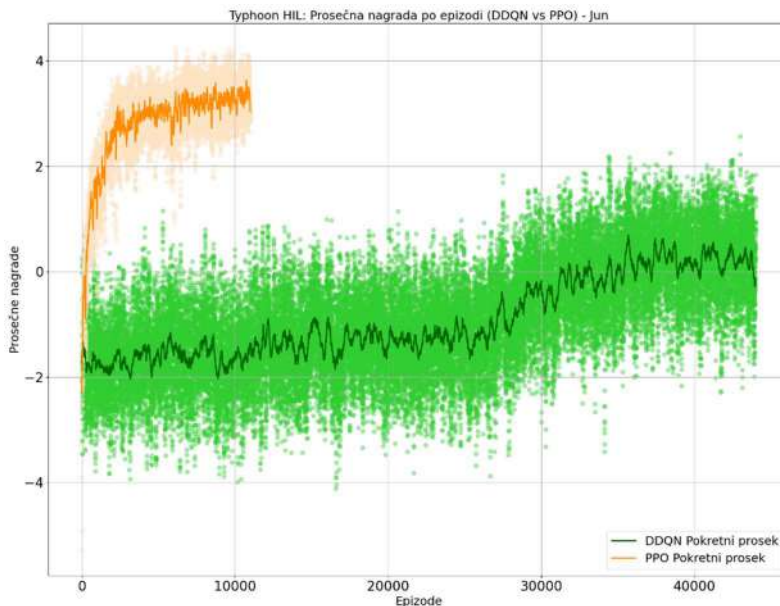


(a) *Typhoon HIL* simulaciono okruženje

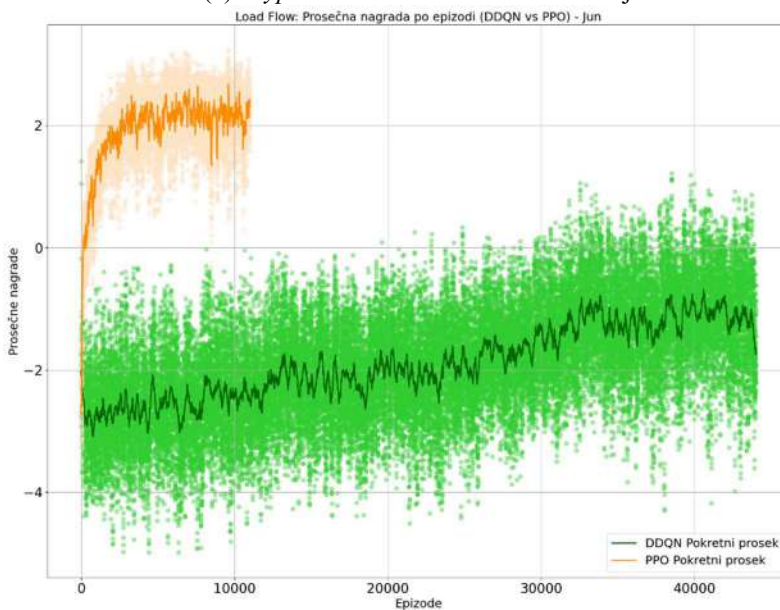


(b) *Load Flow* simulaciono okruženje

Slika 20. Prosečna nagrada po epizodi za zimski mesec (1. do 22. januar).



(a) *Typhoon HIL* simulaciono okruženje



(b) *Load Flow* simulaciono okruženje

Slika 21. Prosečna nagrada po epizodi za letnji mesec (1. do 22. jun).

Rezultati ukazuju na ograničenu konvergenciju DDQN modela. Iako deluje da je DDQN dostigao konvergenciju nagrade, naredne analize su pokazale da ovaj model nije uspešno obučen:

- Analiza odluka modela na testnom danu koji sledi neposredno nakon treniranja je otkrila da je model propustio da uključi već mašinu u *Load Flow* simulacionom okruženju u zimskom danu (potpoglavlje 6.1.3.1)

- Interpretacija odluka koje model donosi (potpoglavlje 6.1.4.1) je otkrila da odluke DDQN modela u zimskom i letnjem danu deluju gotovo nasumično i često ne prate logiku energetske optimizacije i udobnosti stanara.
- Ispitivanje robusnosti modela njegovom primenom na sedam testnih dana koji slede neposredno nakon treniranja je ukazalo na to da model često propušta da upali veš mašinu (potpoglavlje 6.4).

Uzrok ovakvog ishoda leži u činjenici da algoritmi optimizacije zasnovani na gradijentima, koji nemaju formalne garancije konvergencije, mogu dovesti do učenja suboptimalnih politika čak i kada je obezbeđena dovoljna eksploracija (Zhang et al. 2018). Optimizacija u učenju potkrepljivanjem značajno je složenija nego u nagledanom učenju, što dovodi do nestabilnih krivih nagrade i nepredvidivih ishoda treniranja (Zhang et al. 2018). Pored toga, epizodne nagrade tokom obučavanja često osciluju (Mnih et al. 2013), a male promene u inicijalizaciji težina mogu značajno uticati na krajnje performanse (Henderson et al. 2018). Zbog toga se konvergencija krive nagrade ne može tumačiti kao garancija uspešnog i ispravnog obučavanja DDQN modela i neophodno je vršiti dodatne analize ponašanja modela.

6.1.2 Analiza vrednosti funkcije nagrade na testnom danu koji sledi neposredno nakon perioda treniranja

U ovom potpoglavlju sprovedena je analiza vrednosti nagrade koju modeli ostvaruju u danu koji neposredno sledi nakon perioda obuke, sa ciljem ispitivanja njihove sposobnosti da generalizuju naučenu politiku i donose odluke u nepoznatim uslovima. Posebno su razmatrane vrednosti nagrade povezane sa udobnošću stanara, cenom električne energije i njihovim zbirnim efektom.

Rezultati analize biće razdvojeni u zavisnosti od simulacionog okruženja (*Typhoon HIL* i *Load Flow*), pri čemu se za svako okruženje prikazuju rezultati ostvareni u zimskom i letnjem danu. Uređaji simulirani u *Typhoon HIL* okruženju troše manje energije u poređenju sa uređajima simuliranim u *Load Flow* okruženju, zbog čega se vrednosti nagrade ostvarene u različitim okruženjima ne mogu direktno porediti.

U tabeli 16 prikazane su vrednosti nagrade, izražene u evrima, koje su modeli ostvarili na testnim danima (23. januara i 23. juna) u okviru *Load Flow* simulacionog okruženja. Iako deluje da je DDQN u januarskom danu ostvario veću uštedu električne energije u odnosu na PPO model, ovaj rezultat je posledica činjenice da DDQN model nije aktivirao veš mašinu (tabela 16 i slika 24). PPO model je u analiziranom januarskom danu ostvario ukupnu funkciju nagrade veću za 4,22 € u odnosu na DDQN, što predstavlja unapređenje od približno 160%. U junskom testnom danu, oba modela su uspela da organizuju rad uređaja tako da proizvodnja električne energije premašuje potrošnju. PPO je u tom danu ostvario ukupnu funkciju nagrade veću za 0,29 €, što odgovara unapređenju od oko 16% u odnosu na DDQN. Na osnovu rezultata dobijenih za oba testna dana može se zaključiti da je PPO model u *Load Flow* simulacionom okruženju ostvario bolje performanse u pogledu ukupne funkcije nagrade u odnosu na DDQN model. Ovaj rezultat je u skladu sa očekivanjima, budući da je tokom faze obuke PPO model ostvario približno 2 puta veću konačnu nagradu u zimskom, odnosno 2,5 puta veću u letnjem danu u poređenju sa DDQN modelom (potpoglavlje 6.1.1).

Tabela 16. *Load Flow* okruženje: vrednosti nagrade ostvarene na testnom danu (23. dan januara i juna). Negativne vrednosti označavaju da je potrošnja energije bila veća od proizvodnje, dok pozitivne vrednosti označavaju da je proizvodnja električne energije veća od potrošnje.

	Model	Udobnost (27) sa kaznama za nepoštovanje želja stanara [€]	Cena električne energije (32) [€]	Zbir komponenti udobnosti i cene [€]
<i>Januar</i>	DDQN	0,49	-3,11	-2,62
	PPO	5,80	-4,20	1,60
<i>Jun</i>	DDQN	1,09	0,72	1,81
	PPO	1,39	0,67	2,10

Tabela 17 prikazuje performanse modela za 23. dan u januaru i junu u okviru *Typhoon HIL* simulacionog okruženja. U ovom okruženju, DDQN model je pokazao bolje performanse od PPO modela i u januarskom i u junskom danu, posmatrano prema ukupnoj funkciji nagrade. DDQN model je ostvario veću vrednost ukupne funkcije nagrade za 0,06 € (3.6% unapređenja) u januarskom danu i 0.14 € (4.7% unapređenja) u junskom danu u odnosu na PPO model. Ovakvi rezultati nisu u potpunosti u skladu sa vrednostima konačne funkcije nagrade koje su modeli ostvarili tokom faze obuke (potpoglavlje 6.1.1). Moguće objašnjenje za ovaj ishod leži u činjenici da su posmatrani dani posebno povoljni za DDQN model. Naime, analiza robusnosti modela sprovedena na periodu od sedam uzastopnih dana nakon završetka treninga (potpoglavlje 6.4) pokazuje da, iako PPO model ostvaruje slabije rezultate u konkretnim danima januara i juna, on u proseku ostvaruje bolje performanse od DDQN modela u *Typhoon HIL* simulacionom okruženju, posmatrano na sedam testnih dana, za oba perioda godine. I u ovom slučaju, u junskom danu oba modela uspešno su organizovala rad uređaja tako da je ukupna potrošnja električne energije bila manja od proizvedene.

Tabela 17. *Typhoon HIL* okruženje: vrednosti nagrade ostvarene na testnom danu (23. dan januara i juna). Negativne vrednosti označavaju da je potrošnja energije bila veća od proizvodnje, dok pozitivne vrednosti označavaju da je proizvodnja električne energije veća od potrošnje.

	Model	Udobnost (27) sa kaznama za nepoštovanje želja stanara [€]	Cena električne energije (32) [€]	Zbir komponenti udobnosti i cene [€]
<i>Januar</i>	DDQN	5,30	-3,63	1,67
	PPO	5,22	-3,61	1,61
<i>Jun</i>	DDQN	1,36	1,62	2,98
	PPO	1,41	1,43	2,84

Iako PPO model u okviru *Typhoon HIL* simulacionog okruženja ostvaruje nešto slabije rezultate u poređenju sa DDQN modelom (smanjenje od 3,6%, odnosno 4,7%), ovo pogoršanje nije značajno u odnosu na unapređenje koje PPO model pokazuje u *Load Flow* simulacionom okruženju (povećanje od 160% i 16%). Na osnovu toga može se zaključiti da su ukupne performanse PPO modela robusnije u odnosu na različite

postavke simulacionog okruženja. U *Typhoon HIL* simulacionom okruženju, uređaji su trošili više električne energije, pri čemu je njihova potrošnja modelovana kao konstantna. Nasuprot tome, u *Load Flow* simulacionom okruženju, uređaji su trošili manje električne energije, a njihova potrošnja modelovana je kao varijabilna.

U odnosu na konačnu vrednost funkcije nagrade ostvarene tokom faze obuke (potpoglavlje 6.1.1), u okviru *Load Flow* simulacionog okruženja DDQN model je na testnom danu u januaru ostvario ukupnu nagradu manju za 0,62 €, dok je na testnom danu u junu ostvario ukupnu nagradu veću za približno 1,67 €. U okviru *Typhoon HIL* simulacionog okruženja, DDQN model je pokazao izraženija odstupanja u odnosu na nagrade ostvarene tokom obuke – ukupna funkcija nagrade bila je značajno viša, i to za 3,76 € u januarskom i 2,98 € u junskom danu. Ova značajna odstupanja ukazuju na nestabilnost i ograničenu sposobnost generalizacije odluka DDQN modela na dane koje nisu bili obuhvaćeni procesom obuke, što je potvrđeno i analizom robusnosti modela sprovedenom njihovom primenom na sedam testnih dana (potpoglavlje 6.4).

Sa druge strane, ukupna funkcija nagrade koju je PPO model ostvario na testnim danima pokazuje visok stepen konzistentnosti u odnosu na konačne vrednosti nagrade ostvarene tokom tokom faze obuke (potpoglavlje 6.1.1). U okviru *Load Flow* simulacionog okruženja, u januarskom danu PPO model je ostvario ukupnu nagradu za 0.4 € nižu u odnosu na nagradu postignutu tokom obuke, dok je u junskom danu postigao gotovo identičnu ukupnu nagradu, od približno 2 €. U *Typhoon HIL* simulacionom okruženju, PPO model je u januarskom danu ostvario ukupnu nagradu nižu za 0,39 €, dok je u junskom danu ostvario nagradu veću za 0,16 € u odnosu na vrednosti dobijene tokom obuke. Konzistentnost nagrada ostvarenih na testnim danima u odnosu na nagrade postignute tokom obuke ukazuje na to da PPO model poseduje bolju sposobnost generalizacije odluka u poređenju sa DDQN modelom.

Naredno potpoglavlje analizira odluke koje su modeli donosili u testnim danima analiziranim u ovom poglavlju.

6.1.3 Analiza odluka modela na nivou dana

Svrha analize odluke modela na nivou jednog dana, iz perspektive ljudskog eksperta, je da se proceni da li su odluke modela usklađene sa željama stanara (potpoglavlje 5.1.3) uz racionalnu potrošnju električne energije. Analiza je razdvojena po uređajima i obuhvata kontrolu unutrašnje temperature, rad mašine za pranje sudova i rad mašine za veš. Ove analize su grupisane na nivou godišnjeg doba gde potpoglavlje 6.1.3.1 analizira rad uređaja u zimskom danu, a potpoglavlje 6.1.3.2 analizira rad uređaja u letnjem danu.

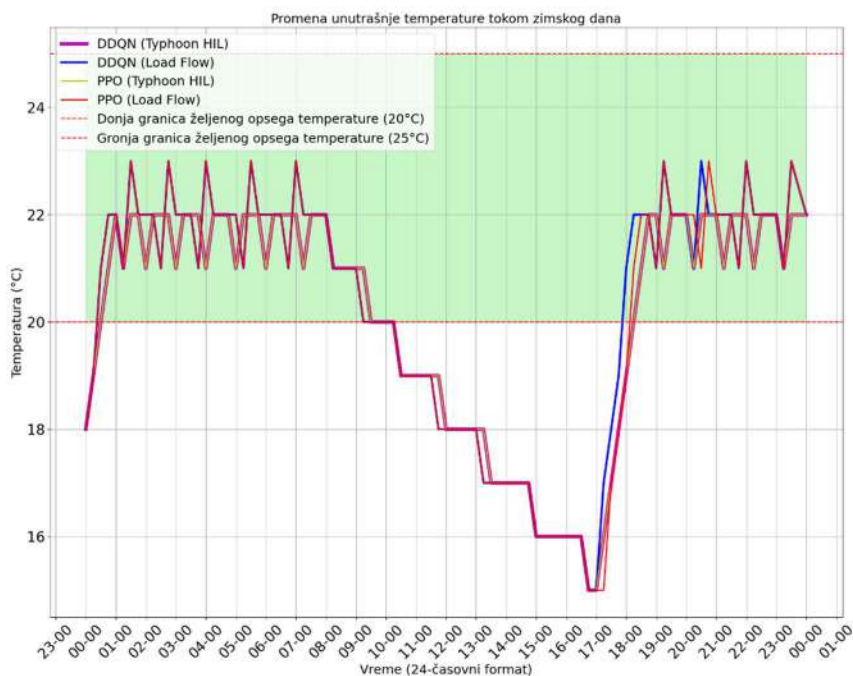
Za evaluaciju modela učenja potkrepljivanjem, odabrani su dani koji nisu korišćeni tokom procesa treniranja modela. Za obučavanje modela korišćena su prva 22 dana u mesecu, dok je za evaluaciju korišćen 23. dan. Analiziran je zimski i letnji dan s ciljem demonstracije zavisnosti odbranih akcija modela od sezonskih uslova.

6.1.3.1 Analiza rada uređaja u zimskom danu

U zimskom danu analiziran je rad tri uređaja pod kontrolom obučanih modela: klima uređaj, mašina za pranje sudova i veš mašina, sa ciljem da se proceni usklađenost njihovih odluka sa spoljašnjim faktorima karakterističnim za ovo doba godine, kao i sa

preferencijama stanara. Pošto DDQN model u okviru *Load Flow* simulacionog okruženja tokom zimskog testnog dana nije aktivirao veš mašinu, u ovom potpoglavlju nije prikazana njegova kontrola tog uređaja.

Slika 22 prikazuje način na koji su DDQN i PPO modeli kontrolisali unutrašnju temperaturu tokom zimskog dana. U oba simulaciona okruženja, PPO model održava unutrašnju temperaturu na sličan način – klima uređaj se isključuje oko 8:00 i uključuje oko 17:00 (u skladu sa željama stanara iz tabele 13), nakon čega se unutrašnja temperatura brzo vraća u željeni opseg. DDQN model u *Typhoon HIL* okruženju održava temperaturu na isti način. Iako DDQN model u *Load Flow* simulacionom okruženju dostigne željeni opseg pre ostalih modela, ne postiže bolje rezultate u pogledu ukupne nagrade jer ne aktivira sve uređaje, što negativno utiče na udobnost stanara.

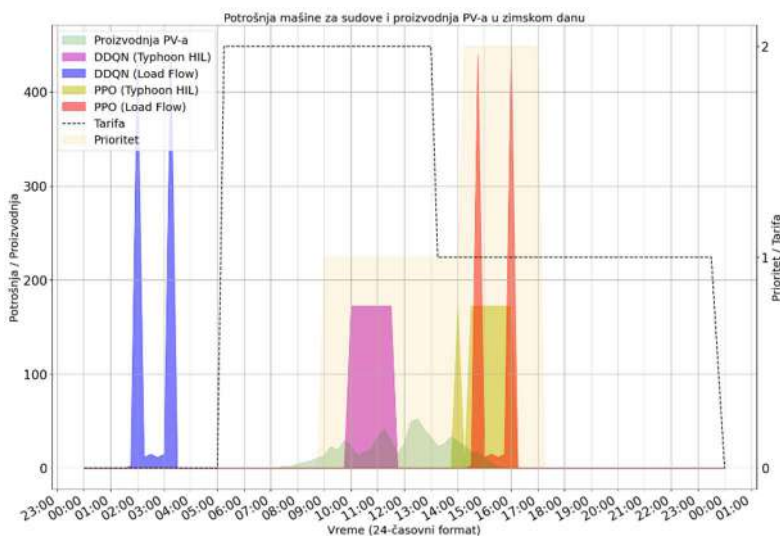


Slika 22. Kontrolisanje unutrašnje temperature u toku zimskog dana.

Slika 23 prikazuje upravljanje mašinom za pranje sudova tokom zimskog dana. PPO model u *Typhoon HIL* simulacionom okruženju ne koristi optimalno mašinu za pranje sudova (nakon pokretanja) jer je ciklus pranja prekinut pre završetka, čime je prethodno uložena energija izgubljena, a model prinuđen da ponovo pokrene proces pranja od početka. Što se tiče perioda pokretanja mašine za pranje sudova, PPO modeli u oba simulaciona okruženja biraju da pokrenu uređaj tokom perioda srednje tarife, kada započinje interval visokog značaja rada ovog uređaja.

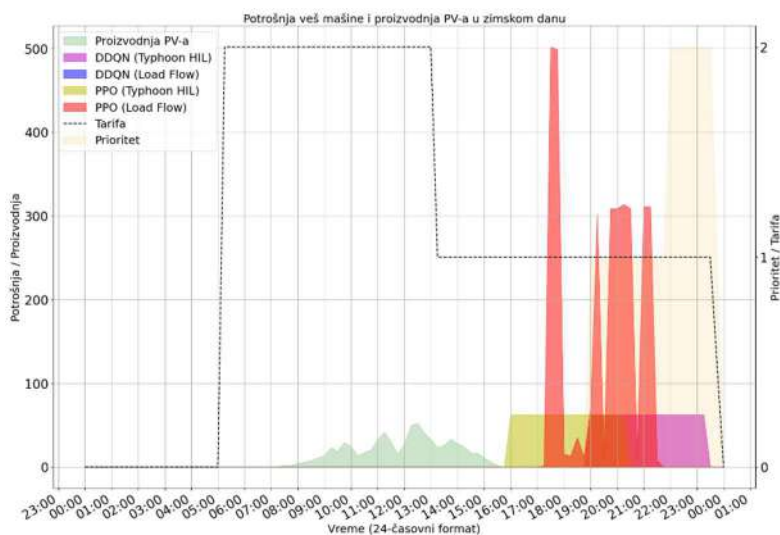
U *Typhoon HIL* simulacionom okruženju, DDQN model aktivira mašinu za pranje sudova u periodu proizvodnje električne energije, pri visokoj tarifi i srednjem značaju rada uređaja. Takvom odlukom, DDQN model, prema rezultatima u tabeli 17, u poređenju sa PPO modelom ostvaruje bolji balans između udobnosti stanara i uštede

električne energije. DDQN model u *Load Flow* simulacionom okruženju pokreće mašinu dosta ranije u odnosu na ostale modele, u periodu niske tarife.



Slika 23. Upravljanje mašinom za sudove u zimskom danu.

Slika 24 prikazuje način upravljanja veš mašinom tokom zimskog dana. PPO modeli aktiviraju uređaj pre početka perioda srednjeg značaja rada, i to u vreme trajanja srednje tarife električne energije. PPO model u *Typhoon HIL* simulacionom okruženju pokreće uređaj znatno pre nego što započne željeni opseg rada. Sa druge strane, DDQN model u *Typhoon HIL* simulacionom okruženju aktivira veš mašinu u periodu srednje tarife, tokom srednjeg i visokog značaja rada uređaja. U poređenju sa PPO modelom u istom simulacionom okruženju, DDQN model je ostvario bolje rezultate u pogledu ukupne nagrade, što je potvrđeno i rezultatima prikazanim u tabeli 17. DDQN model je u *Load Flow* simulacionom okruženju propustio da aktivira mašinu za veš.



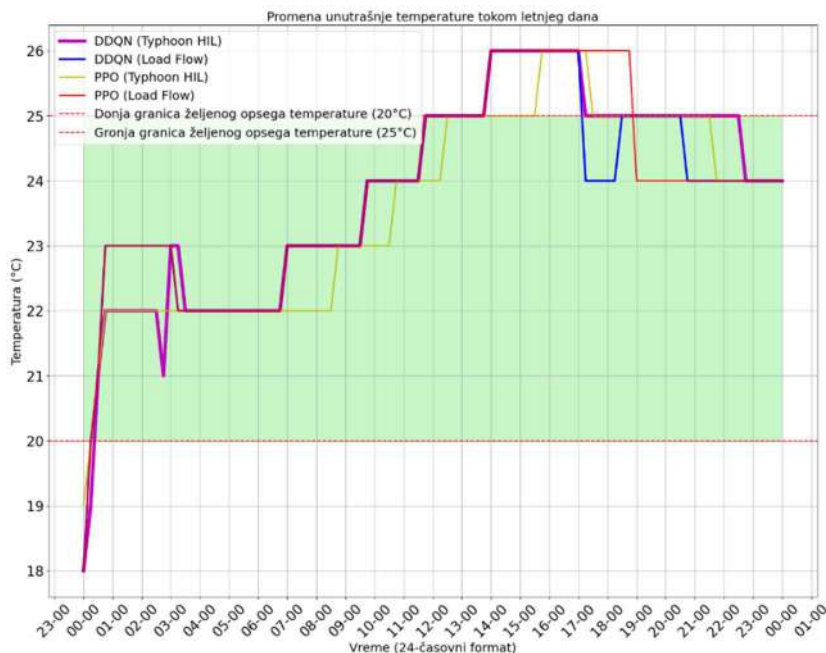
Slika 24. Upravljanje mašinom za veš u zimskom danu.

Razlika u upravljanjem radom mašine za pranje sudova i mašinom za veš ogleda se u tome što su modeli pridali veću važnost periodu značaja rada mašine za pranje sudova, dok period značaja rada mašine za veš nije imao podjednak uticaj na proces donošenja odluka. Odluke DDQN modela u *Typhoon HIL* simulacionom okruženju pokazale su se boljim u odnosu na odluke PPO modela. U *Load Flow* simulacionom okruženju, DDQN model nije uspešno pokrenuo mašinu za veš, što je rezultovalo nižom cenom električne energije, ali i smanjenom udobnošću stanara.

6.1.3.2 Analiza rada uređaja u letnjem danu

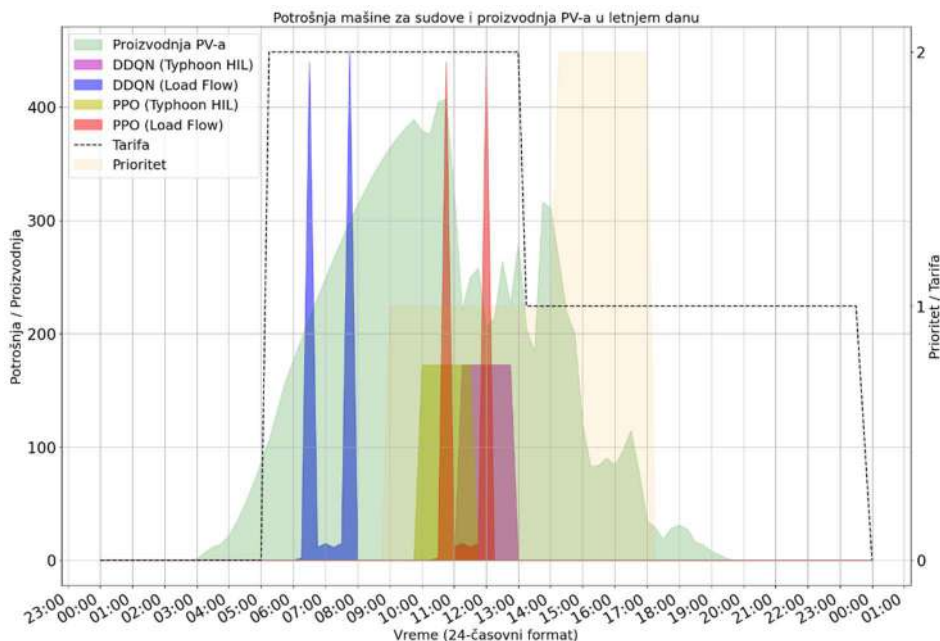
Tokom letnjeg dana posmatran je rad klima uređaja, mašine za pranje sudova i mašine za veš pod kontrolom obučenih modela. Analiziran je uticaj spoljašnjih faktora karakterističnih za letnje uslove na odluke koje modeli donose.

Slika 25 prikazuje regulaciju unutrašnje temperature tokom letnjeg dana. Temperatura kod PPO modela na kratko izlazi iz željenog temperaturnog opsega – u *Load Flow* simulacionom okruženju željeni temperaturni opseg dostiže se tek od 19 časova. Ostali modeli upravljaju radom klima uređaja na sličan način, ali uspevaju da dostignu željeni temperaturni opseg od 17 časova, u skladu sa željama stanara. U *Load Flow* simulacionom okruženju PPO model ostvaruje bolje rezultate u pogledu uštede električne energije, jer u poređenju sa DDQN modelom ređe aktivira klima uređaj (tabela 16). Takođe, postignuti su nešto bolji rezultati i u pogledu udobnosti, iako se ova prednost PPO modela u odnosu na DDQN ne ogleda u načinu upravljanja klima uređajem. Sa druge strane, u *Typhoon HIL* simulacionom okruženju, DDQN postiže bolje rezultate u poređenju sa PPO modelom (tabela 17), ali se o ova prednost ne može u potpunosti uočiti samo na osnovu analize upravljanja klima uređajem.



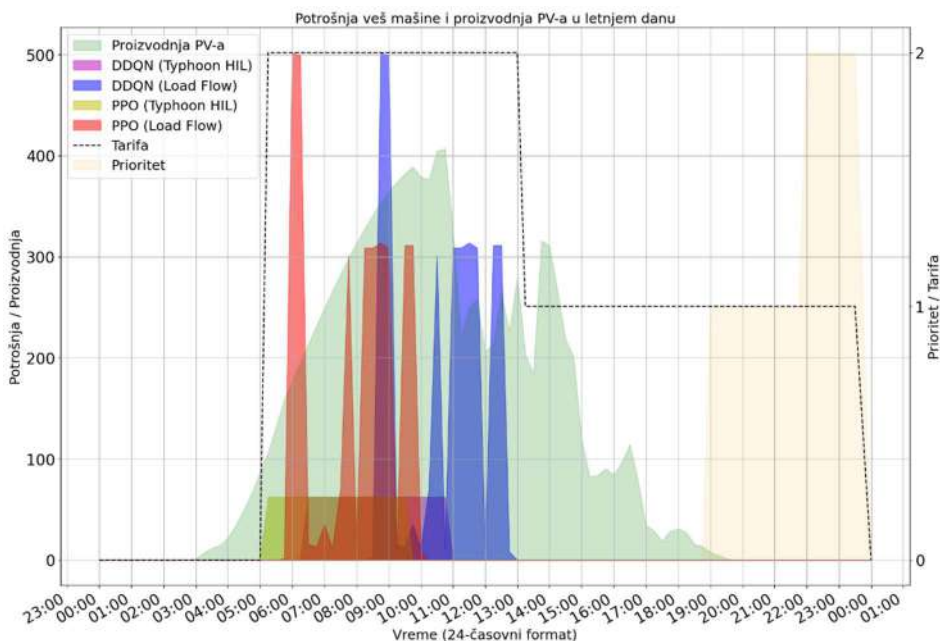
Slika 25. Kontrolisanje unutrašnje temperature u toku letnjeg dana.

Na slici 26 prikazan je način na koji modeli upravljaju radom mašine za pranje sudova tokom letnjeg dana. Svi modeli aktiviraju mašinu za pranje sudova u periodu kada solarni paneli proizvode energiju. DDQN model u *Load Flow* simulacionom okruženju pokreće mašinu za pranje sudova pre početka perioda srednjeg značaja rada i u vreme povećane proizvodnje solarne energije. Zbog toga, u poređenju sa PPO modelom u istom okruženju, ostvaruje slabije rezultate u pogledu udobnosti stanara. PPO model u *Load Flow* simulacionom okruženju aktivira mašinu za pranje sudove u periodu srednjeg značaja rada. U *Typhoon HIL* simulacionom okruženju, vreme aktiviranja mašine je gotovo identično kod oba modela, pa se na osnovu toga ne može jasno sagledati od čega potiče prednost DDQN modela u odnosu na PPO (tabela 17).



Slika 26. Upravlja mašinom za sudove u letnjem danu.

Na slici 27 prikazana je kontrola rada mašine za veš tokom letnjeg dana. Algoritmi pokreću mašinu za veš u vremenskom intervalu koji je sličan periodu aktiviranja mašine za sudove. Svi modeli započinju rad uređaja tokom perioda visoke tarife, ali istovremeno i u vreme povećane proizvodnje električne energije iz solarnog panela. Nijedan od modela ne uzima u obzir želje stanara vezane za vreme rada uređaja, što ukazuje na to da je prioritet stavljen na optimizaciju potrošnje električne energije nauštrb udobnosti stanara.



Slika 27. Upravljanje mašinom za veš u letnjem danu.

6.1.3.3 Zaključak analize odluka model na nivou zimskog i letnjeg dana

Na osnovu prethodno prikazanih rezultata (Slike 23–27) može se uočiti da se logika ponašanja modela razlikuje između zimskog i letnjeg dana usled promene spoljašnjih uslova i dostupnosti izvora energije. Tokom zimskog dana, modeli su usmereni na održavanje komfora stanara, jer je proizvodnja energije putem solarnog panela smanjena. U tom periodu modeli teže postizanju ravnoteže između potrošnje i udobnosti, pri čemu je ključan faktor tarifni režim električne energije. Nasuprot tome, u letnjem danu prisustvo solarne proizvodnje značajno menja prioritete modela. U letnjem danu odluke se donose sa ciljem maksimalnog korišćenja raspoložive solarne energije, čak i po cenu smanjenja komfora stanara. Time se fokus modela pomera sa balansiranja između tarifa na optimizaciju potrošnje energije, što objašnjava različite obrasce ponašanja u letnjem danu.

Rezultati ukazuju da logika donošenja odluka o aktiviranju uređaja nije trivijalna niti ujednačena tokom cele godine. Ona zavisi od kombinacije sezonskih faktora, tarifnih perioda i definisanih opsega značaja rada uređaja. Ručno podešavanje rada uređaja zahteva definisanje posebnih obrazaca upravljanja u zavisnosti od doba godine, pri čemu je neophodno istovremeno uzeti u obzir više međuzavisnih parametara, što dodatno komplikuje proces. Značaj automatizovanog sistema za upravljanje uređajima u pametnoj kući ogleda se u njegovoj sposobnosti da integriše različite spoljne i unutrašnje faktore i donosi optimalne odluke, koje bi stanarima bile složene i zahtevne za svakodnevno planiranje.

6.1.4 Analiza uticaja karakteristika stanja na odluke modela

U ovom potpoglavlju biće analizirane odluke modela, gde će se za modele zasnovane na učenju potkrepljivanjem sprovesti detaljna analiza odlučivanja optimalnih akcija u posmatranim stanjima korišćenjem SHAP biblioteke. Vizuelna interpretacija SHAP vrednosti realizovana je putem dijagrama sažetka značaja atributa (eng. *summary plot*) i dijagrama sile doprinosa (eng. *force plot*), koji omogućavaju analizu uticaja karakteristika stanja na odluke modela. U potpoglavlju 6.1.4.1 interpretiraju se odluke na osnovu više uzoraka stanja, i interpretiraju se odluke u određenim koracima simulacije. Analize su grupisane na nivou simulacionih okruženja i dalje podeljenje na osnovu godišnjeg doba. Za interpretacije modela na globalnom i lokalnom nivou korišćeno je sedam testnih dana u januaru i junu.

6.1.4.1 Interpretacija odluka

Interpretacija odluka može da se izvrši na osnovu više uzoraka stanja tj. na globalnom nivou i u određenom vremenskom koraku tj. na lokalnom nivou.

Interpretacija na globalnom nivou se zasniva na dijagramu sažetka značaja koji prikazuju agregirane SHAP vrednosti za svaku karakteristiku stanja kroz više uzoraka stanja. Konkretno, za svaku moguću akciju aktivacije uređaja, prikupljeni su svi uzorci stanja S_t u toku posmatranog perioda (sedam testnih dana). Za svako stanje S_t , SHAP vrednost $\phi_i(S_t)$ je izračunata za svaku karakteristiku stanja i , i ova vrednost pokazuje koliko bi se odluka modela da izvrši akciju promenila ako bi ta karakteristika bila uklonjena iz ulaza.

Na y-osi prikazane su karakteristike stanja, rangirane prema prosečnom apsolutnom doprinosu odluci modela (odabranoj akciji), pri čemu se najuticajnije nalaze na vrhu. Sortiranje se vrši na osnovu jednačine:

$$\text{MEAN}(|\phi_i|) = \frac{1}{N} \sum_{t=1}^T |\phi_i(S_t)|,$$

gde T predstavlja broj uzoraka u evaluaciji (sedam testnih dana po sezoni, gde svaki dan ima 96 vremenskih koraka t zbog 15 minutnog koraka simulacije).

X-osa prikazuje SHAP vrednosti $\phi_i(S_t)$ koje kvantifikuju doprinos konkretne vrednosti karakteristike na akciju koju je model odabrao. Svaka tačka na dijagramu predstavlja SHAP vrednost izračunatu za jednu karakteristiku u konkretnom uzorku stanja. Boja tačke ukazuje na vrednost te karakteristike u posmatranom stanju: crvena označava visoke, a plava niske vrednosti. Položaj tačke duž horizontalne ose pokazuje da li ta vrednost pozitivno ili negativno utiče na odabranu akciju u datom stanju, i koliki je intenzitet tog uticaja. Rasipanje tačaka ilustruje kako se doprinos određene karakteristike menja u različitim stanjima, tj. koliko je njen uticaj stabilan ili zavisen od konteksta. Na osnovu dijagrama moguće je zaključiti koje karakteristike stanja najviše doprinose odabranim akcijama, kako se njihov uticaj menja u zavisnosti od konkretnih vrednosti karakteristike, i da li postoji konzistentan pravac tog uticaja. Na taj način se dobija globalna slika o tome koje informacije model koristi pri izboru akcija.

Dijagram sile doprinosa koristi se za lokalnu interpretaciju, odnosno za objašnjenje zašto je model odabrao određenu akciju u konkretnom stanju. Za razliku od globalne interpretacije koja prikazuje opšti uticaj karakteristika na ponašanje modela, lokalna

analiza se fokusira na jednu specifičnu predikciju i doprinos svake karakteristike toj odluci. Vizualizacija se zasniva na pomeranju vrednosti predikcije u odnosu na referentnu vrednost (na primer, prosečnu predikciju za sve uzorke eng. *base value*).

$$\text{base value} = \mathbb{E}[f(S_t)],$$

gde je $f(S_t)$ predikcija modela. SHAP vrednosti za svaku karakteristiku u posmatranom stanju se zatim sabiraju sa baznom vrednošću:

$$f(S_t) = \text{base value} + \sum_{i=1}^M \phi_i(S_t),$$

gde je M broj karakteristika stanja. Ovim se dobija ukupna predikcija modela za stanje S_t .

Svaka karakteristika stanja prikazana je pravougaonikom duž horizontalne ose, pri čemu crvena boja označava pozitivan doprinos (povećava verovatnoću izbora date akcije), a plava negativan doprinos (smanjuje verovatnoću izbora date akcije). Karakteristike sa jačim uticajem nalaze se bliže centralnoj granici između pozitivnih i negativnih doprinosa, dok širina pravougaonika reflektuje intenzitet njihovog efekta $|\phi_i(S_t)|$. Ovakav prikaz omogućava intuitivno razumevanje ponašanja agenta u konkretnoj situaciji, posebno u stanjima koja odstupaju od tipičnih obrazaca. U kombinaciji sa globalnom interpretacijom, lokalna interpretacija doprinosi dubljem razumevanju ponašanja modela, kako na nivou pojedinačnih odluka, tako i u kontekstu ukupne strategije učenja potkrepljivanjem.

Izdvojićemo nekoliko primera kako bismo ilustrovali razloge za odluke modela u određenim stanjima, koristeći dijagram sile doprinosa. Ova analiza omogućuje bolje razumevanje uticaja pojedinačnih faktora na odluke modela u različitim simulacionim okruženjima.

6.1.4.1.1 Odluke u zimskom periodu

Slike 28 i 29 prikazuju agregirane srednje apsolutne SHAP vrednosti za sve akcije, sa ciljem da se identifikuju koje karakteristike stanja najviše utiču na odluke modela.

Slika 28 ukazuje na to da PPO model ima konzistentan obrazac značaja karakteristika u oba simulaciona okruženja. Najveći uticaj na predikcije modela ima proizvedena energija iz PV, što ukazuje da PPO efikasno koristi raspoloživu solarnu energiju prilikom planiranja aktivacije uređaja. Sledeće bitne karakteristike su značaj rad klima uređaja, stanja mašine za sudove i mašine za veš i termalni komfor korisnika izražen kao razlika između unutrašnje i željene temperature. Redosled najuticajnijih karakteristika na predikciju modela je isti u *Load Flow* i *Typhoon HIL* simulacionim okruženjima. Ova stabilnost u ponašanju sugerise na robusnost u donošenju odluka od strane PPO modela, što ukazuje na sposobnost generalizacije u različitim simulacionim okruženjima.



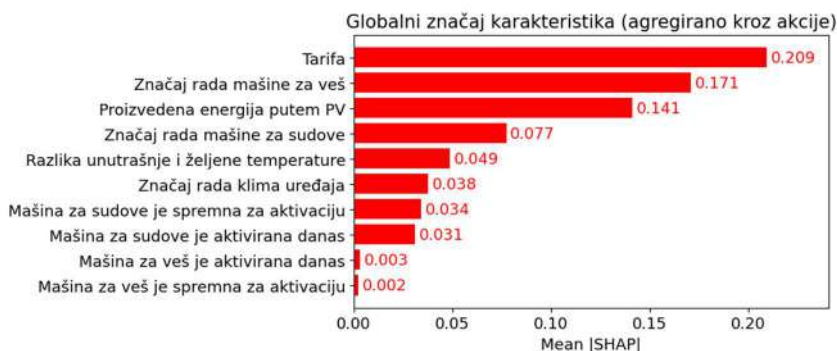
(a)



(b)

Slika 28. PPO model: agregiran uticaj karakteristika stanja kroz akcije u toku zimskog perioda.

Slika 29 ukazuje na to da DDQN model ispoljava drugačiju logiku ponašanja u odnosu na PPO model. Na predikcije DDQN modela najveći uticaj ima tarifa električne energije, nakon čega slede značaj rada veš mašine i proizvodnja energije putem solarnog panela. Takođe treba istaći da, za razliku od PPO modela, rang značaja karakteristika kod DDQN modela se razlikuje u različitim simulacionim okruženjima. U oba okruženja karakteristike vezane za mašinu za veš imaju zanemarljiv uticaj na konačnu odluku modela.



(a)



(b)

Slika 29. DDQN model: agregiran uticaj karakteristika stanja kroz akcije tokom zimskog perioda.

Slike 28 i 29 pružaju globalni uvid u ponašanje modela, ali ovaj pristup ima određena ograničenja. Srednja apsolutna vrednost prikriva vremensku dinamiku, te prividno visoke vrednosti određene karakteristike mogu proisteći iz nekoliko izolovanih momenata, a ne iz doslednog uticaja tokom celog perioda. Takođe, pošto se SHAP vrednosti agregiraju preko svih akcija, ne može se jasno sagledati doprinos svake pojedinačne karakteristike na konkretne odluke modela.

Na slici 30 prikazan je uticaj karakteristika stanja na odluku o uključivanju klima uređaja tokom zimskog dana za PPO model. PPO modeli su u oba simulaciona okruženja uspešno održavali unutrašnju temperaturu u željenom opsegu tokom visokog značaja rada klima uređaja (slika 22). Na slici 30 vidimo da su odluke modela zasnovane na karakteristikama stanja koje su zaista relevantne za ovu odluku. Konkretno, karakteristike stanja sa najvećim doprinosom odluci o uključivanju klima uređaja u oba simulaciona okruženja su sledeće:

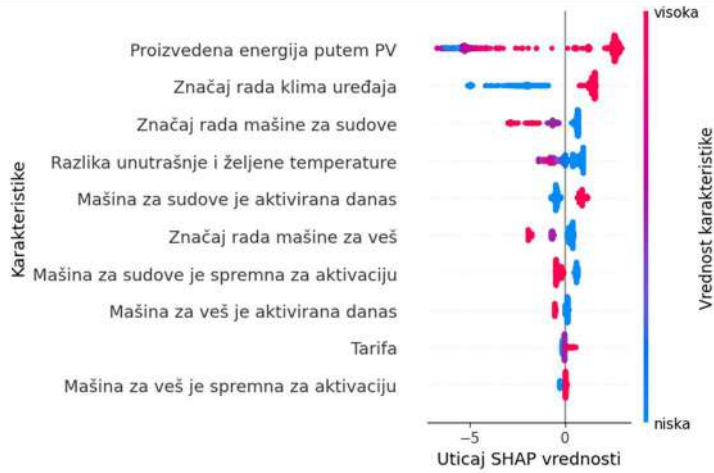
- niska proizvodnje energije putem solarnog panela (S_{PV}),
- visok značaj rada klima uređaja (S_P^{AC}),
- nizak značaj rada mašine za sudove (S_P^{DW}) i
- visok temperaturni komfor, odnosno niska razlika unutrašnje i željene temperature ($S_{\Delta T_{razlika}}$).

Važno je naglasiti da velike negativne vrednosti na x -označavaju veliku PV proizvodnju, dok vrednosti bliske 0 znače da je PV proizvodnja niska. Takođe, razlika unutrašnje i željene temperature je apsolutna vrednosti, gde niske vrednosti označavaju manje odstupanje unutrašnje temperature od željene i obrnuto, odnosno, visoke vrednosti označavaju veliko odstupanje.

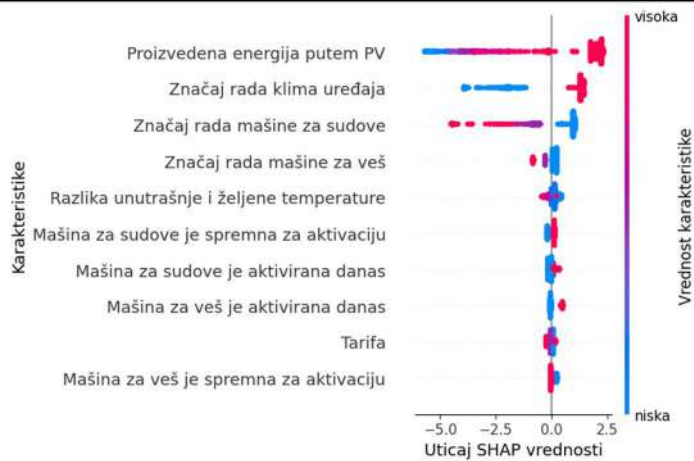
Na slici 30 vidimo da ponašanje modela deluje kontraintuitivno u odnosu na karakteristiku proizvodnje energije iz solarnog panela. Pri visokoj PV proizvodnji, SHAP vrednosti su negativne, odnosno model smanjuje verovatnoću izbora akcije uključivanja klima uređaja kada je raspoloživa veća količina proizvedene energije. Takođe, u zoni niske PV proizvodnje, pojavljuje se mali broj slučajeva sa pozitivnim SHAP doprinosom, što znači da model u tim instancama procenjuje da je uključivanje klima uređaja optimalna akcija. Ovakvo ponašanje može delovati kontraintuitivno, jer niska proizvodnja energije intuitivno sugerise manju spremnost sistema da uključi dodatno opterećenje.

Na prvi pogled, nelogično ponašanje se ogleda u pozitivnom doprinosu odluci o aktiviranju klima uređaja za niske vrednosti apsolutne vrednosti temperaturne razlike (veći komfor) i negativnom doprinosu odluci o aktiviranju klima uređaja u većim apsolutnim vrednostima razlike (niži komfor) u *Typhoon HIL* simulacionom okruženju. Niske apsolutne vrednosti temperaturne razlike ($1^{\circ}\text{C} - 2^{\circ}\text{C}$) označavaju visok nivo komfora, zato što je unutrašnja temperatura u željenom opsegu, što rezultuje pozitivnom doprinosu odluci da se klima uređaj aktivira. Suprotno tome, visoke apsolutne vrednosti ukazuju na veća odstupanja od željene temperature, odnosno manji komfor, čime negativno utiču na odluku da se klima uređaj aktivira.

Iako takvo ponašanje deluje kontraintuitivno, moguće objašnjenje je u načinu na koji je značaj rada klima uređaja definisan od strane stanara (tabela 13), odnosno stanarima nije bitan rad klime u periodu njihovog odsustva, te aktiviranje klima uređaja od strane treniranog PPO modela u toku dana, kada stanari nisu prisutni, ne rezultuje pozitivnom nagradom. Dok je održavanje temperature, odnosno komfora stanara, rezultovalo znatno većom nagradom. Većina pozitivnih odluka o aktiviranju klima uređaja doneta je u uslovima visokog značaja rada klima uređaja, male temperaturne razlike i niske PV proizvodnje. Suprotno tome, situacije sa visokom temperaturnom razlikom, koje takođe mogu rezultovati pozitivnom nagradom, javljaju se retko, pretežno oko 17:00 časova, kada je značaj rada uređaja najviši (tabela 13). Budući da su takvi slučajevi brojčano zanemarljivi u odnosu na dominantne, model nije razvio stabilan obrazac koji bi prepoznao veliku temperaturnu razliku kao pozitivan faktor, što se odrazilo i u SHAP analizi. Dodatni uticaj niskog značaja rada mašine za sudove može se objasniti definisanim značajem rada uređaja od strane stanara (tabela 13). Naime, u periodima kada se klima uređaj aktivirao, značaj rada mašine za sudove bio je nizak.

Typhoon HIL

(a)

Load Flow

(b)

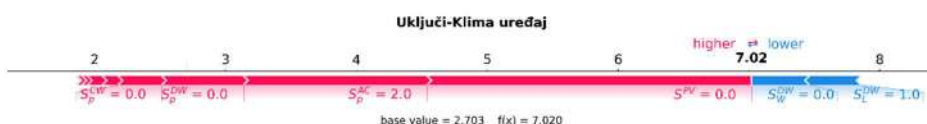
Slika 30. PPO model: uticaj karakteristika stanja na akciju “*uključi klima uređaj*” tokom zimskog perioda.

Iako se SHAP vrednosti često koriste za interpretaciju ponašanja RL modela, njihova analiza u kompleksnim okruženjima može dovesti do kontraintuitivnih rezultata. Naime, smer uticaja pojedinih karakteristika na odluke modela ne mora nužno odgovarati očekivanjima ili realnom doprinosu tih karakteristika performansama agenta. Jedan od ključnih razloga za to je što SHAP pretpostavlja nezavisnost karakteristika i odsustvo međusobne korelacije (Engelhardt et al. 2024) (Salih et al. 2025). Zbog te pretpostavke, analiza zavisnosti između pojedinačnih SHAP vrednosti i odluka modela može biti ograničene pouzdanosti, jer u realnim RL sistemima odluke često proizilaze iz složene interakcije više osobina. Upravo ta nemogućnost SHAP-a da obuhvati zajedničke efekte i nelinearne zavisnosti između karakteristika predstavlja značajan nedostatak njegove primene u kompleksnim RL sistemima (Engelhardt et al. 2024).

SHAP predstavlja post-hoc metod interpretacije odluka koji omogućavaju identifikaciju osobina sa najvećim uticajem na odluke treniranog model (Merasha et al. 2024). Posebno

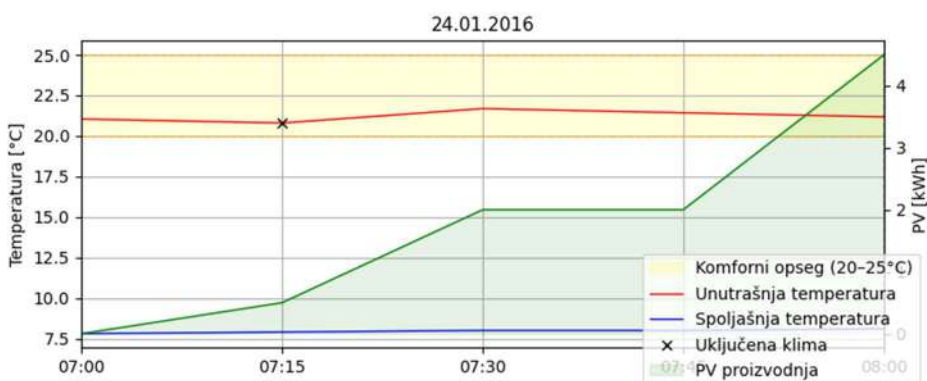
je koristan za poređenje značaja pojedinačnih osobina u okviru različitih modela ili scenarija. Rezultate SHAP analize neophodno je prikazivati zajedno sa odgovarajućim grafikovima, čime se olakšava razumevanje doprinosa svake osobine izlazu modela i pretpostavki na kojima se metoda zasniva (Salih et al. 2025).

Na slici 31 je prikazana odluka PPO modela da “uključi klima uređaj” u zimskom periodu u *Typhoon HIL* simulacionom okruženju u jednom vremenskom koraku (između 7:00 i 8:00) Pozitivan uticaj na odluku o aktiviranju klima uređaja imaju niska proizvodnja energije ($S^{PV} = 0$), nizak značaj rada mašine za sudove ($S_P^{DW} = 0.0$) i mašine za veš (S_P^{CW}). Ovde je uticaj proizvodnje energije kontrainuitivan, dok je očekivano da pozitivan uticaj imaju niski značaj rada ostalih uređaja. Nelogičan pozitivan uticaja niske vrednosti temperaturne razlike, uočen na globalnom nivou (slika 30), nije imao značajan uticaj na odluku o aktiviranju klima uređaja u ovom vremenskom koraku (slika 31). Slika 32 prikazuje kombinaciju faktora koja odgovara opisanoj situaciji. Konkretno, grafikom se prikazuje vremenski trenutak označen simbolom x, koji predstavlja momenat aktiviranja klima uređaja. Na slici se vidi da model u ovom slučaju aktivira klima uređaj dok se temperatura kreće u željenom temperaturnom opsegu kako bi obezbedio da unutrašnja temperatura ostane unutar opsega dok je visok značaj rada klima uređaja (do 8:00). Stoga, iako smer uticaja niske proizvodnje energije deluje kontra-intuitivno, vidi se da odluka sistema jeste intuitivna kada se u obzir uzme kombinacija više različitih faktora.



Slika 31. PPO model: uticaj karakteristika na akciju “uključi klima uređaj” u *Typhoon HIL* simulacionom okruženju u zimskom danu za stanje u vremenskom koraku 07:15.

Temperatura i PV proizvodnja (07:00–08:00)



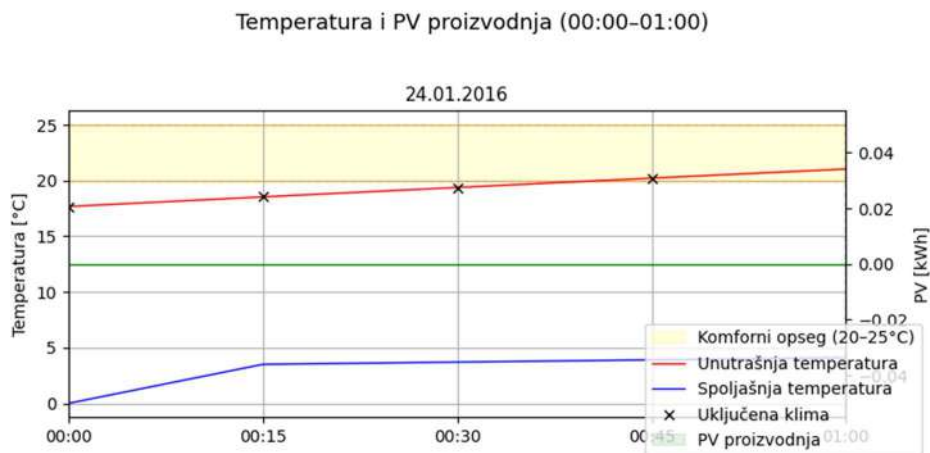
Slika 32. PPO model: upravljanje unutrašnjom temperaturom uz pregled proizvodnje energije putem solarnog panela i temperaturni komfor između 7:00 i 8:00 sati, u *Typhoon HIL* simulacionom okruženju.

Na slici 33 je prikazana odluka PPO modela da “uključi klima uređaj” u zimskom periodu u *Typhoon HIL* simulacionom okruženju na početku dana, gde smer uticaja nije logičan

za karakteristike S^{PV} (proizvodnja energije putem solarnog panela) i $S_{\Delta T_{razlika}}$ (razlika unutrašnje i željene temperature). Pozitivan uticaj na odluku o aktiviranju klima uređaja ima niska proizvodnja energije, dok negativan uticaj ima velika temperaturna razlika (niska temperaturna udobnost). Uvid u ponašanje PPO modela početkom dana je prikazan na slici 34, gde grafikom prikazuje trenutak aktiviranje klima uređaja označen simbolom x u 00:15 sati. U tom trenutku, unutrašnja temperatura nalazi se van željenog temperaturnog opsega, dok nema proizvodnje energije od strane PV-a. Ovakva kombinacija karakteristika u potpunosti odgovara vrednostima prikazanim na slici 33, što potvrđuje da je model odluku o uključivanju klima uređaja doneo u situaciji niske PV proizvodnje i visoke temperaturne razlike, sa ciljem da obezbedi stabilnost unutrašnje temperature tokom perioda visokog značaja rada uređaja. Iako je kontrainuitvan negativan uticaj visoke temperaturne razlike, na slici 22 se vidi da je model uspešno naučio da održava temperaturu u željenom temperaturnom opsegu.



Slika 33. PPO model: uticaj karakteristika na akciju “uključi klima uređaj” u *Typhoon HIL* simulacionom okruženju u zimskom danu za stanje u vremenskom koraku 00:00.



Slika 34. PPO model: upravljanje unutrašnjom temperaturom uz pregled proizvodnje energije putem solarnog panela i temperaturni komfor između 00:00 i 01:00 sati, u *Typhoon HIL* simulacionom okruženju.

Na slici 35 prikazan je uticaj karakteristika stanja na odluku o uključivanju klima uređaja tokom zimskog dana za DDQN model. DDQN modeli u oba simulaciona okruženja su uspešno održavali unutrašnju temperaturu u željenom opsegu tokom viskog značaja rada klima uređaja (slika 22). Karakteristike stanja sa najvećim doprinosom odluci o uključivanju klima uređaja su sledeće:

- niske vrednosti tarife (tarifa),

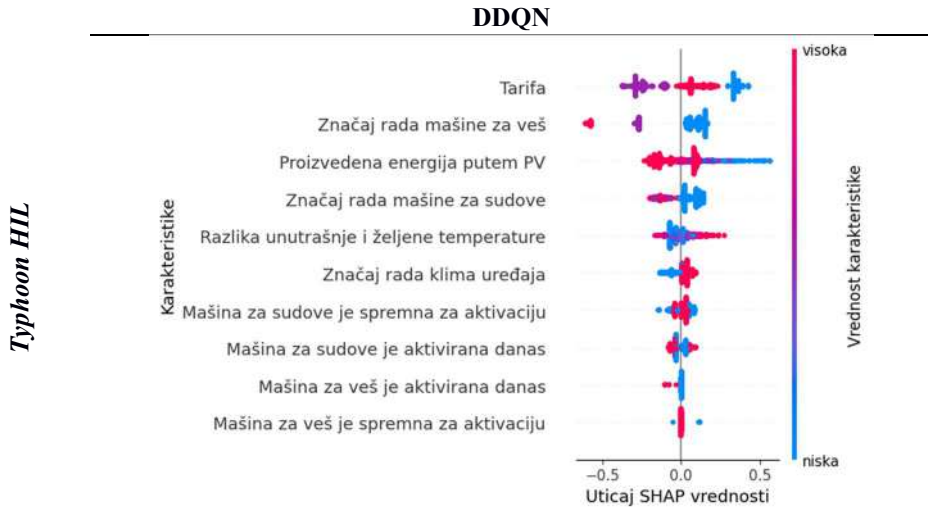
- visoka PV proizvodnja u *Typhoon HIL* simulacionom okruženju i kombinacija visoke i niske PV proizvodnje u *Load Flow* simulacionom okruženju (S_{PV}),
- niži značaja rada mašine za sudove u *Typhoon HIL* simulacionom okruženju (S_P^{DW}) i niži značaja rada mašine za veš u oba simulaciona okruženja (S_P^{CW}),
- viši značaj rada klima uređaja u *Typhoon HIL* simulacionom okruženju i niži i viši značaj rada u *Load Flow* simulacionom okruženju (S_P^{AC}),
- manji termalni komfor, odnosno više vrednosti razlike unutrašnje i željene temperature $S_{\Delta T_{razlika}}$.

Za razliku od PPO modela, kod DDQN modela u oba simulaciona okruženja tarifa ima značajnu ulogu u donošenju odluke o radu klima uređaja.

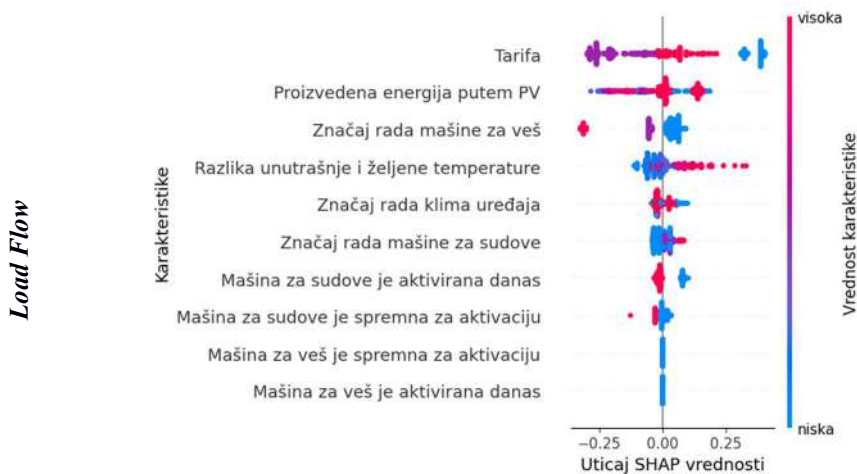
Kod DDQN modela u *Load Flow* simulacionom okruženju, gde model nije bio uspešno obučen da aktivira veš mašinu, takođe se vidi da na odluku o aktiviranju klima uređaja imaju nelogičan uticaj sledeće karakteristike: niska PV proizvodnja (S_{PV}) i nizak značaj rada klima uređaja (S_P^{AC}).

Pozitivan uticaj niske PV proizvodnje na odluku o aktivaciji klima uređaja se može objasniti na isti način kao u prethodnim primerima, gde se period visokog značaja rada klima uređaja (tabela 13) poklapao sa periodom niske PV proizvodnje. Sa druge strane, period visoke PV proizvodnje se u zimskom periodu retko poklapao sa visokim značajem rada klima uređaja, s obzirom na to da je PV proizvodnja bila generalno niska u toku dana.

Dodatni uticaj niskog značaja rada mašina za sudove i veš može se objasniti definisanim značajem rada uređaja od strane stanara (tabela 13). Naime, u periodima kada se klima uređaj aktivirao, značaj rada mašina za sudove i veš bili su niski.



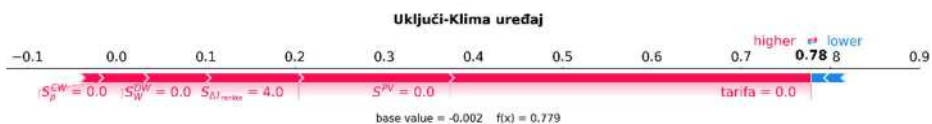
(a)



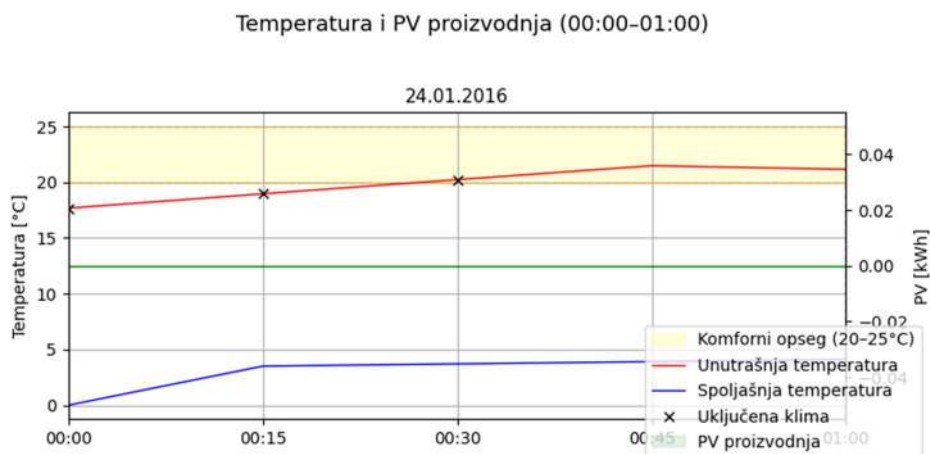
(b)

Slika 35. DDQN model: uticaj karakteristika stanja na akciju “uključi klima uređaj” tokom zimskog perioda.

Na slici 36 prikazana su karakteristike koje su imale uticaj na odluku o aktiviranju klima uređaja od strane DDQN modela u *Load Flow* simulacionom okruženju za akciju izvršenu u vremenskom trenutku 00:00. Pozitivan uticaj na odluku imaju karakteristike: niska tarifa, visoke vrednosti temperaturne razlike i (kontraintuitivno) niska PV proizvodnja. Slika 37 prikazuje aktiviranje klima uređaja u 00:00, označen simbolom x, gde su prethodno pomenute karakteristike imale uticaj na odluku. Unutrašnja temperatura se kreće ka željenom temperaturnom opsegu. Odsustvo i umanjeno značaja rada klima uređaja na odluku modela može da proizilazi iz činjenice da tarifa ima jak uticaj na odluku DDQN modela (slika 29).

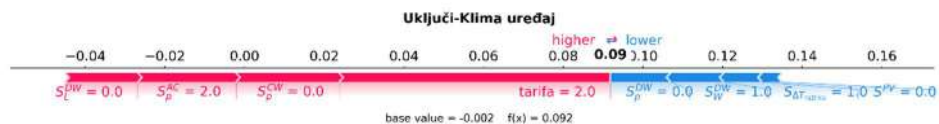


Slika 36. DDQN model: uticaj karakteristika na akciju “uključi klima uređaj” u *Load Flow* simulacionom okruženju u zimskom danu za stanje u vremenskom koraku 00:00.



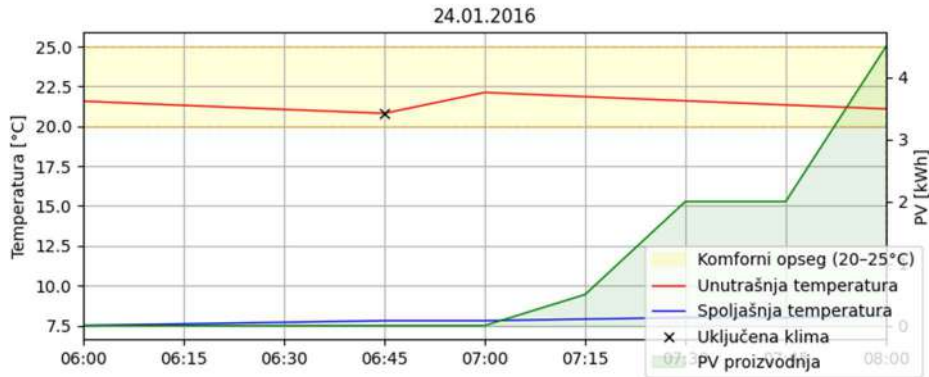
Slika 37. DDQN model: upravljanje unutrašnjom temperaturom uz pregled proizvodnje energije putem solarnog panela i temperaturni komfor između 00:00 i 01:00 sati, u *Load Flow* simulacionom okruženju.

Na slici 38 prikazan je uticaj karakteristika na aktiviranje klima uređaja od strane DDQN model u *Load Flow* simulacionom okruženju u vremenskom koraku 06:45. Konkretni vremenski korak je prikazan na slici 39 simbolom x u 06:45. Karakteristike koje su imale pozitivan uticaj na odluku o uključivanju klima uređaja su: (kontraintuitivno) visoka tarifa (*tarifa*), nizak značaj rada mašine za već (S_p^{CW}), visok značaj rada klima uređaja (S_p^{AC}) i činjenica da mašina za sudove nije spremna za rad (S_L^{DW}). Ova kombinacija karakteristika odgovara upravljanju klimom uređaja prikazanom na slici 39. Pozitivan uticaj visoke tarife je nelogičan i proizilazi iz treniranog DDQN modela koji je pridao veliki značaj tarifi, dok je zanemario ostale karakteristike (slika 29).



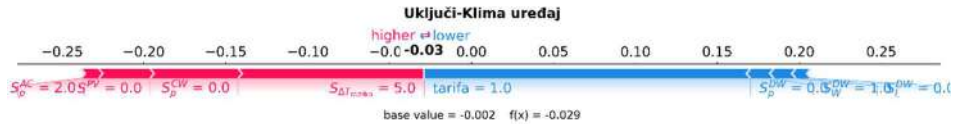
Slika 38. DDQN model: uticaj karakteristika na akciju “uključi klima uređaj” u *Load Flow* simulacionom okruženju u zimskom danu za stanje u vremenskom koraku 06:45.

Temperatura i PV proizvodnja (06:00–08:00)

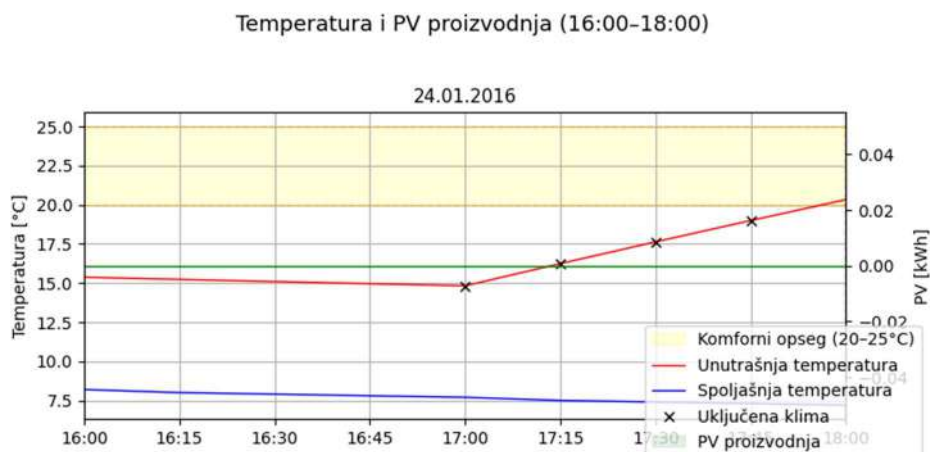


Slika 39. DDQN model: upravljanje unutrašnjom temperaturom uz pregled proizvodnje energije putem solarnog panela i temperaturni komfor između 06:00 i 08:00 sati, u *Load Flow* simulacionom okruženju.

Slika 40, prikazuje vremenski korak 17:15, koji je takođe na slici 41 označen simbolom x. U ovom slučaju su karakteristike sa pozitivnim uticajem na odluku uključivanja klima uređaja bile visoka temperaturna razlika ($S_{\Delta T_{\text{razlika}}}$), nizak značaj rada mašine za već (S_p^{CW}), niska PV proizvodnja (S^{PV}) i visok značaj rada klima uređaja (S_p^{AC}). Sa druge strane, veliki negativan uticaj je imala srednja vrednost tarife. Neintuitivna je činjenica da je visok značaja rada klima uređaja imao slab uticaj na odluku, što se ogleda i u redosledu prioriteta karakteristika kod treniranog DDQN modela (slika 29).



Slika 40. DDQN model: uticaj karakteristika na akciju “uključi klima uređaj” u *Load Flow* simulacionom okruženju u zimskom danu za stanje u vremenskom koraku 17:15.

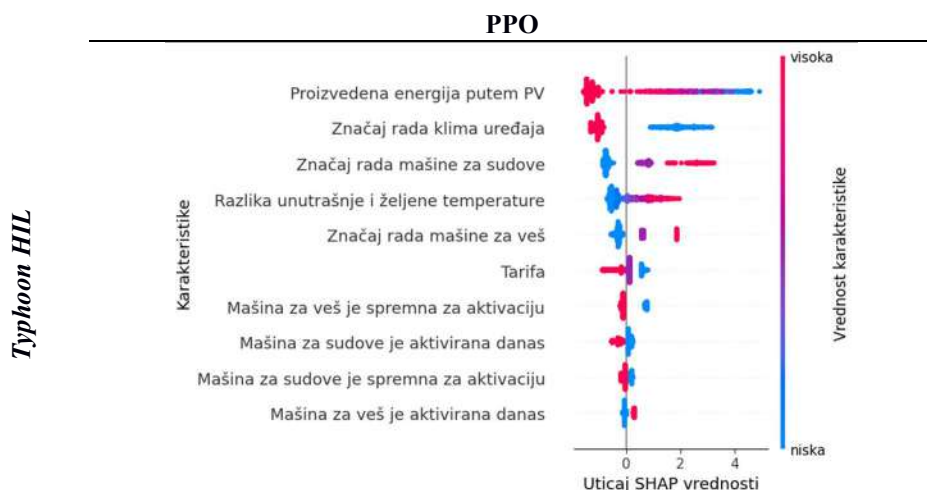


Slika 41. DDQN model: upravljanje unutrašnjom temperaturom uz pregled proizvodnje energije putem solarnog panela i temperaturni komfor između 16:00 i 18:00 sati, u *Load Flow* simulacionom okruženju.

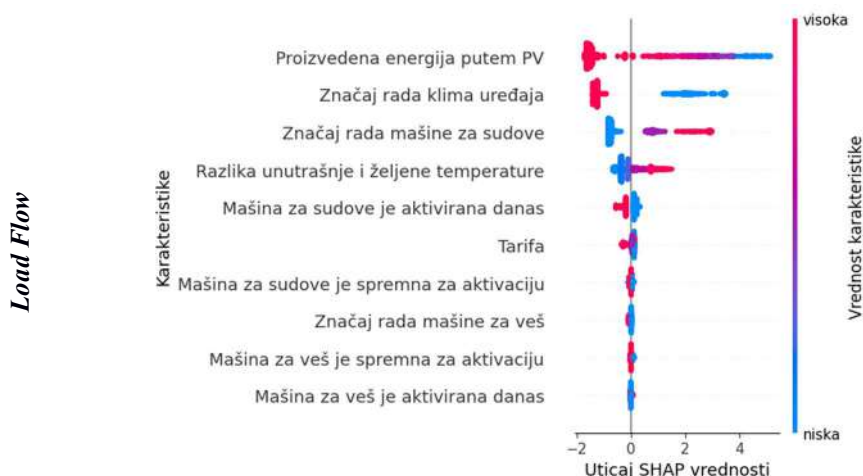
Na slici 42 prikazane su karakteristike koje utiču na akciju aktiviranja mašine za sudove kod PPO modela. Karakteristike koje imaju uticaj na odluku u oba simulaciona okruženja su:

- visoka proizvodnja električne energije S_{PV} ,
- nizak značaj rada klima uređaja S_p^{AC} ,
- visok i srednji značaj rada mašine za sudove S_p^{DW} .

Na osnovu slike 23 uočava se jasno poklapanje između pozitivnog uticaja navedenih karakteristika i stvarnih odluka PPO modela o aktiviranju mašine za sudove. Model je mašinu aktivirao u trenucima visoke proizvodnje električne energije, kada je značaj rada klima uređaja bio nizak, a značaj mašine za sudove visok, što ukazuje na usklađenost odluka modela sa očekivanim obrascima racionalnog upravljanja energijom.



(a)



(b)

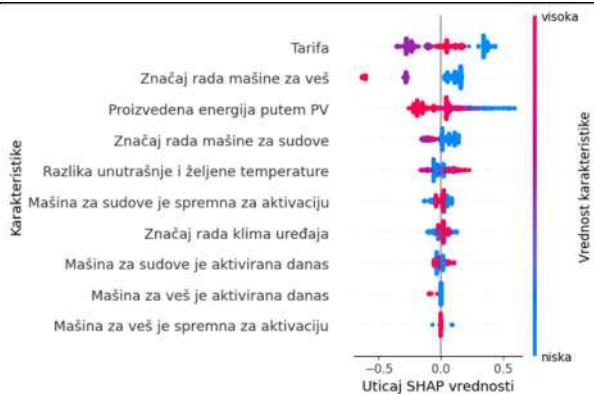
Slika 42. PPO model: uticaj karakteristika stanja na akciju “uključi mašinu za sudove” tokom zimskog perioda.

Na slici 43 prikazane su karakteristike koje utiču na odluku “uključi mašinu za sudove” kod DDQN modela. Karakteristike koje su imali uticaj na odluku u oba simulaciona okruženja su:

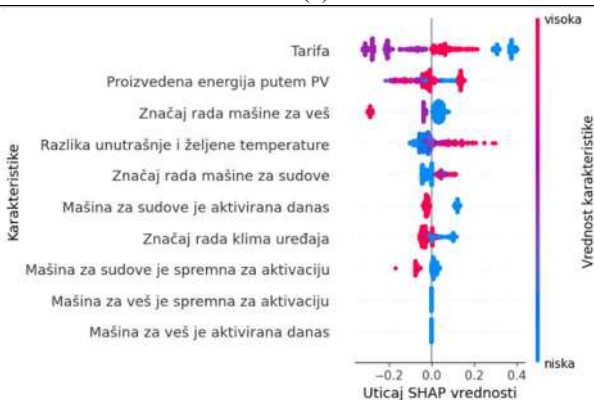
- niska tarifa $tarifa$,
- visoka PV proizvodnja električne energije u *Typhoon HIL* simulacionom okruženju i niska PV proizvodnja električne energije u *Load Flow* simulacionom okruženju S^{PV} ,
- nizak značaj rada mašine sa veš S_p^{CW} ,
- nizak značaj rada mašine za sudove u *Typhoon HIL* simulacionog okruženja i visok značaj rada kod *Load Flow* simulacionog okruženja S_p^{DW} i
- manji termalni komfor tj. visoka razlika unutrašnje i željene temperature $S_{\Delta T_{razlika}}$.

Uočava se nelogičan pozitivan uticaj niskog značaja rada mašine za sudove u *Typhoon HIL* simulacionom okruženju i niske proizvodnje električne energije u *Load Flow* okruženju na odluku modela da aktivira mašinu za sudove. Ovakvo ponašanje odstupa od očekivanog, jer bi aktiviranje mašine za sudove trebalo da bude podstaknuto visokim značajem rada mašine za sudove i povećanom PV proizvodnjom električne energije.

U poređenju sa konkretnim posmatranim danom na slici 23, odluka DDQN modela ne odgovara vrednostima karakteristike značaja rada mašine za sudove. Konkretno u *Typhoon HIL* simulacionom okruženju mašina se aktivirala u periodu srednjeg značaja, dok u *Load Flow* simulacionom okruženju mašina za sudove se aktivira u periodu niskog značaja rada uređaja.

Typhoon HIL

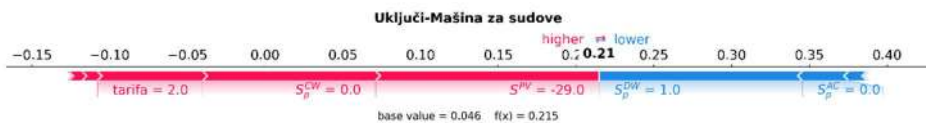
(a)

Load Flow

(b)

Slika 43. DDQN model: uticaj karakteristika stanja na akciju “uključi mašinu za sudove” tokom zimskog perioda.

Slika 44 ukazuje da je DDQN model u *Typhoon HIL* simulacionom okruženju odlučio da uključi mašinu sudove pod uticajem visoke PV proizvodnje električne energije (S^{PV}), niskog značaja rada mašine za veš (S^{CW}) i visoke tarife (*tarifa*). Međutim, činjenica da je model visoku tarifu tretirao kao faktor koji povećava verovatnoću aktiviranje mašine za sudove predstavlja nelogično ponašanje, jer bi visoka tarifa trebalo da ima suprotan efekat odlaganja rada uređaja. Mogući razlog za ovo ponašanje modela je što se period visoke PV proizvodnje preklapa sa periodom visoke tarife. Ako je PV proizvodnja dovoljno visoka da pokrije potrošnju uređaja, ne troši se energija iz mreže, te visoka vrednost tarife nema uticaja na funkciju nagrade. Dodatno, negativan doprinos srednjeg značaja rada mašine za sudove (S^{DW}) je nelogičan. Intuitivno bi bilo da povećan značaj rada uređaja doprinosi većoj verovatnoći njegovog aktiviranja. Sporna karakteristika u *Typhoon HIL* simulacionom okruženju (slika 43 (a)), nizak značaj rada mašine za sudove, nije imala ulogu na aktiviranje uređaja.



Slika 44. DDQN model: Uticaj karakteristika na akciju “uključi mašinu za sudove” u *Typhoon HIL* simulacionom okruženju u zimskom danu za stanje u određenom vremenskom koraku.

Na slici 45 prikazano je da su na odluku DDQN modela o aktiviranju mašine za sudove u *Load Flow* okruženju pozitivno uticale niska tarifa (*tarifa*) i, nelogično, niske vrednosti značaja rada mašine za sudove (S_p^{CW}) i PV proizvodnje S^{PV} . Iako je model reagovao u periodu niske tarife, što je ekonomski opravdano, DDQN u *Load Flow* okruženju nije uspeo da ostvari ravnotežu između uštede energije i udobnosti stanara kao u *Typhoon HIL* okruženju.

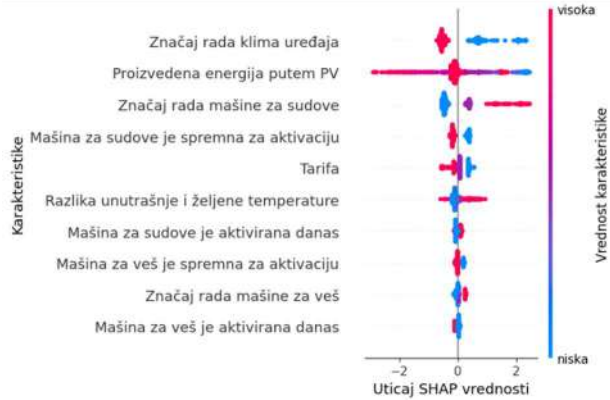


Slika 45. DDQN model: Uticaj karakteristika na akciju “uključi mašinu za sudove” u *Load Flow* simulacionom okruženju u zimskom danu za stanje u određenom vremenskom koraku.

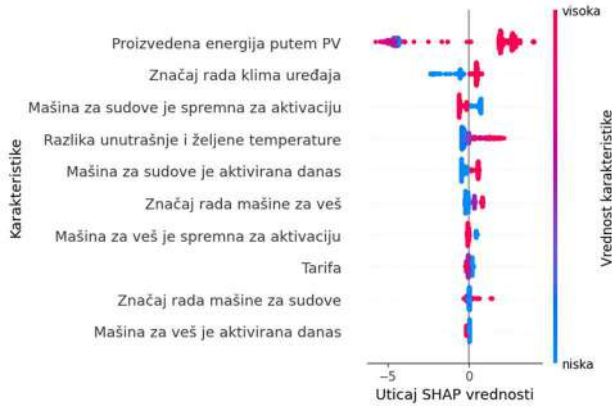
Na slici 46 prikazane su karakteristike koje u PPO modelu utiču na aktiviranje mašine za pranje veša. Karakteristike koje imaju uticaj u oba simulaciona okruženja su:

- visoka PV proizvodnja u *Typhoon HIL* simulacionom okruženju i niska PV proizvodnja električne energije u *Load Flow* simulacionom okruženju (S^{PV}),
- nizak značaj rada klima uređaja u *Typhoon HIL* simulacionom okruženju i visok značaj rada klima uređaja u *Load Flow* simulacionom okruženju (S_p^{AC}) i
- visok značaj rada mašine za sudove je imao jači uticaj u *Typhoon HIL* simulacionom okruženju u odnosu na *Load Flow* okruženje (S_p^{DW}).

U oba okruženja, karakteristike vezane za mašinu za veš nisu imale značajan uticaj na odluku o aktivaciji ovog uređaja. U *Typhoon HIL* simulacionom okruženju, mašina za pranje veša se aktivirala nešto pre 16:00, u periodu visokog značaja rada mašine za pranje sudova i niskog značaja rada klima uređaja (slika 24). U *Load Flow* simulacionom okruženju, su takođe značajne karakteristike se karakteristike poput visokog značaja rada klima uređaja i činjenice da mašina za sudove više nije spremna za rad (pošto je prethodno završila sa radom). Kontraintuitivan je pozitivan uticaj niske PV proizvodnje na odluku aktiviranja mašine za veš, što ukazuje da je treniran PPO model, vođen funkcijom nagrade, u ovom slučaju dao prednost udobnosti u odnosu na uštedu električne energije.

Typhoon HIL

(a)

Load Flow

(b)

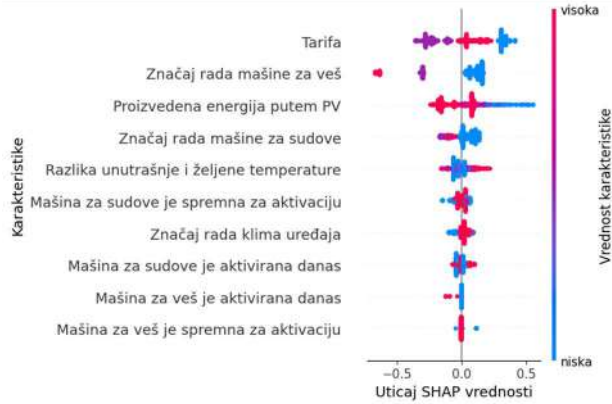
Slika 46. PPO model: uticaj karakteristika stanja na akciju “*uključi veš mašinu*” tokom zimskog perioda.

Na slici 47 analizira se odluka DDQN modela da aktivira veš mašinu u zimskom danu. Karakteristike koje pozitivno utiču na odluku o aktivaciji ovog uređaja su:

- niska tarifa (*tarifa*),
- nizak značaj rada mašine za veš (S_p^{CW}) i
- visoka PV proizvodnja u *Typhoon HIL* simulacionom okruženju i niska PV proizvodnja u *Load Flow* simulacionom okruženju (S^{PV}).

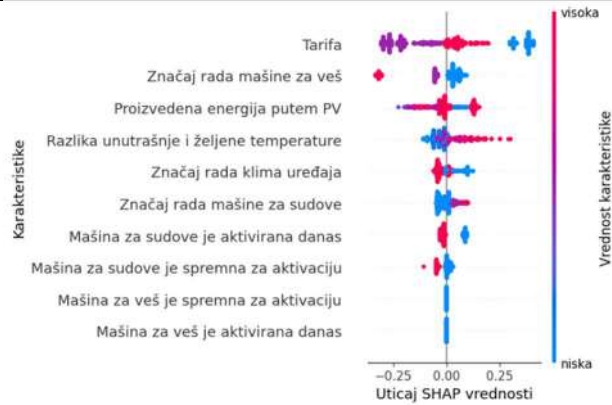
Odluka DDQN modela da aktivira veš mašinu u *Typhoon HIL* simulacionom okruženju (slika 24) u potpunosti se podudara sa karakteristikama koje su identifikovane kao najuticajnije u globalnoj SHAP analizi (slika 46). Naime, model je aktivirao uređaj u periodu niske tarife i neočekivano niskog značaja rada mašine za veš, što su upravo faktori koji prema analizi najviše doprinose odluci o aktivaciji. Jedini faktor koji se ne poklapa je pozitivan uticaj visoke PV proizvodnje. U *Load Flow* simulacionom okruženju, DDQN model je propustio da aktivira mašinu za veš u toku dana.

Typhoon HIL



(a)

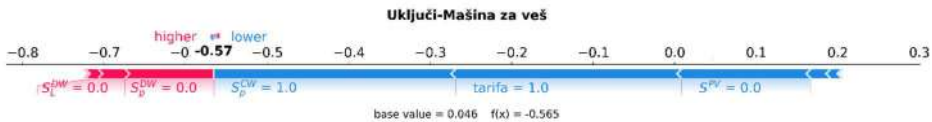
Load Flow



(b)

Slika 47. DDQN model: uticaj karakteristika stanja na akciju “uključi veš mašinu” tokom zimskog perioda.

Na slici 48 prikazano je da obučeni DDQN model pokazuje nelogičan negativan uticaj karakteristike srednjeg značaja rada mašine za veš (S_p^{CW}). Za razliku od globalne analize (slika 47), gde je visoka PV proizvodnja (S^{PV}) imala pozitivan doprinos odluci o aktivaciji, u ovom konkretnom slučaju ta karakteristika nema značajan uticaj na odluku da se mašina za veš uključi.



Slika 48. DDQN model: Uticaj karakteristika na akciju “uključi mašinu za veš” u Typhoon HIL simulacionom okruženju u zimskom danu za stanje u određenom vremenskom koraku.

6.1.4.1.1.1 Zaključak na nivou odluka u zimskom periodu

U zimskom danu modeli su pokazali uticaj različitih karakteristika na donošenje odluka. Odluke PPO modela ukazuju na više-ciljnu optimizaciju, gde postoji dvostruki uticaj

proizvodnje električne energije i značaj rada uređaja, čime se ostvaruje uravnotežen odnos uštede i udobnosti stanara. PPO model je pokazao konzistentnost i stabilnost u oba simulaciona okruženja, pri čemu su odluke bile zasnovane na logičnim energetskim obrascima i realnim uslovima sistema.

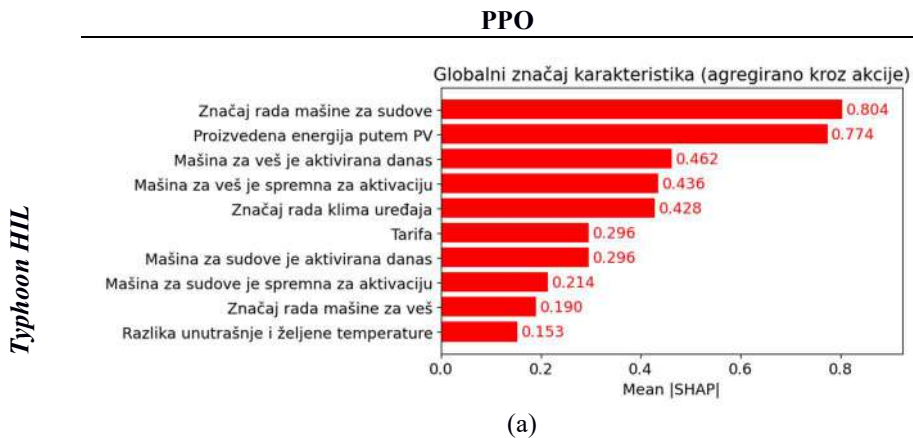
Suprotno tome, DDQN model je pokazao veći stepen osetljivosti na pojedinačne karakteristike, pri čemu je tarifa imala dominantan uticaj na odluke. Takvo ponašanje ukazuje na jednostranu optimizaciju usmerenu ka smanjenju potrošnje električne energije, često na račun udobnosti korisnika. Dodatno, u *Load Flow* simulacionom okruženju uočene su nelogičnosti u smeru uticaja pojedinih karakteristika, što ukazuje na ograničenu sposobnost modela da generalizuje obrasce ponašanja u različitim uslovima.

Generalno, PPO model se pokazao robusnijim i pouzdanijim, sa sposobnošću donošenja odluka koje su u skladu sa racionalnim energetskim principima, dok DDQN model pokazuje ograničenja u interpretabilnosti i stabilnosti, naročito u dinamičnim okruženjima poput *Load Flow* simulacije.

6.1.4.1.2 Odluke u letnjem periodu

Slike 49 i 50 prikazuju agregirane srednje apsolutne SHAP vrednosti za sve akcije sa ciljem da se identifikuju koje karakteristike stanja najviše utiču na odluke modela.

U oba okruženja, najveći uticaj na predikcije modela imaju karakteristike PV proizvodnja, značaj rada klima uređaja i značaj rada mašine za sudove, što ukazuje da PPO efikasno koristi raspoloživu solarnu energiju prilikom planiranja aktivacija uređaja. Redosled značaja ostalih karakteristika je različit u *Typhoon HIL* i *Load Flow* okruženjima. U oba okruženja, razlika unutrašnje i spoljašnje temperature nema veliki uticaj na odluke modela jer, u letnjim danima, spoljašnja temperatura nije značajno različita u odnosu na unutrašnju.



Load Flow



(b)

Slika 49. PPO model: uticaj karakteristika stanja agregirano kroz akcije u toku letnjeg perioda.

Kod DDQN modela (slika 49), rezultati ove analize su različiti. Glavni uticaj na predikcije modela imaju PV proizvodnja i tarifa električne energije. U *Load Flow* simulacionom okruženju, jedino još značaj rada mašine za veš ima značajan uticaj na odluke modela. Kao i kod PPO modela, najniži uticaj ima razlika unutrašnje i željene temperature.

DDQN

Typhoon HIL



(c)

Load Flow



(d)

Slika 50. DDQN model: uticaj karakteristika stanja agregirano kroz akcije tokom letnjeg perioda.

Na slikama 51 i 54 prikazan je uticaj vrednosti karakteristika na odluku o uključivanju klima uređaja tokom letnjeg dana kod PPO i DDQN modela.

Kod PPO modela (slika 51), u oba simulaciona okruženja, sledeće karakteristike imaju značajan uticaj na rad klima uređaja:

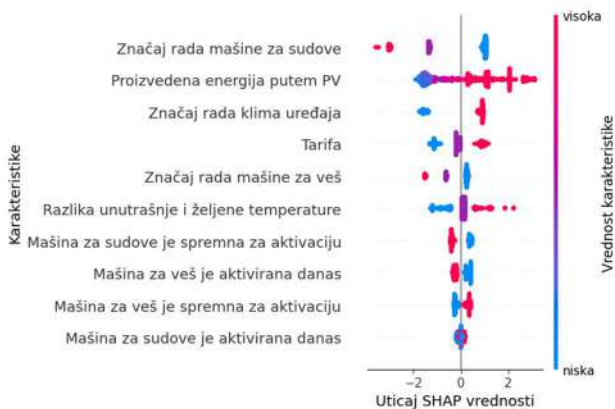
- niži značaj rada mašine za sudove (S_p^{DW}),
- niska PV proizvodnja (S^{PV}),
- visok značaj rada klima uređaja (S_p^{AC}) i
- visoka tarifa.

Na slici 25 prikazana je kontrola rada klima uređaja tokom letnjeg dana, gde se uočava poklapanje navedenih uvida (slika 51) sa ponašanjem PPO modela u oba simulaciona okruženja. Model je dosledno aktivirao klima uređaj u periodima kada je značaj rada ovog uređaja bio visok, a značaj rada ostalih uređaja nizak, čime je obezbeđeno održavanje željenog termalnog komfora korisnika.

Kontraintuitivan pozitivan doprinos niske PV proizvodnje na odluku o aktiviranju klima uređaja može se objasniti definisanim značajem rada uređaja od strane stanara (tabela 13). Tokom perioda odsutnosti stanara, termalna udobnost nije prioritet, pa aktiviranje klima uređaja od strane PPO modela u tim trenucima ne donosi pozitivnu nagradu.

Typhoon HIL

PPO



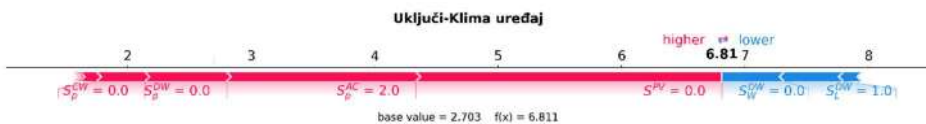
(a)



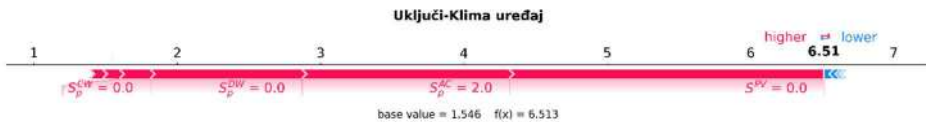
(b)

Slika 51. PPO model: uticaj karakteristika stanja na akciju “uključi klima uređaj” tokom letnjeg perioda.

Na slikama 52 i 53 prikazan je nelogičan pozitivan uticaj niske PV proizvodnje (S^{PV}) u *Typhoon HIL* i *Load Flow* simulacionom okruženju. Iako je na globalnom nivou (slika 51) uočen sporan pozitivan uticaj visoke tarife (*tarifa*), u ovim konkretnim slučajevima ova karakteristika nema značajan uticaj na odluku.



Slika 52. PPO model: Uticaj karakteristika na akciju “uključi klima uređaj” u *Typhoon HIL* simulacionom okruženju u letnjem danu za stanje u određenom vremenskom koraku.



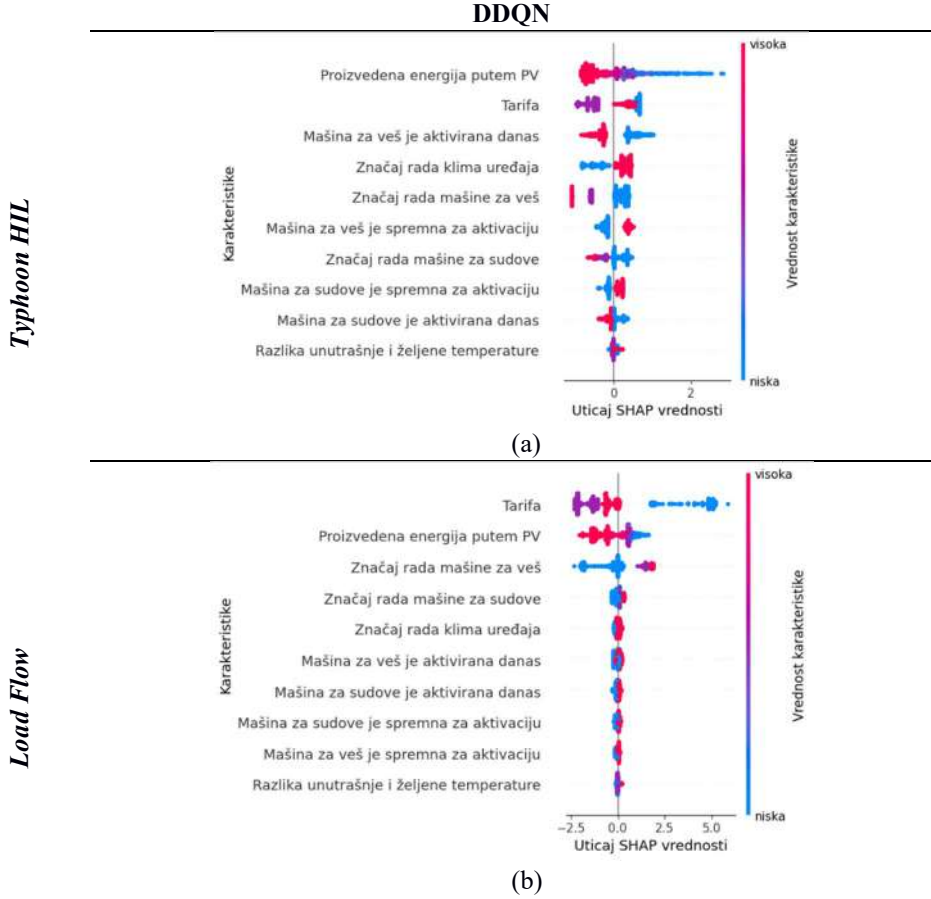
Slika 53. PPO model: Uticaj karakteristika na akciju “uključi klima uređaj” u *Load Flow* simulacionom okruženju u letnjem danu za stanje u određenom vremenskom koraku.

Karakteristike koje imaju uticaj na odluke DDQN model u oba simulaciona okruženja su (slika 54):

- visoka PV proizvodnja (S^{PV}),
- niska tarife,
- karakteristike vezane za mašinu za veš: d u *Typhoon HIL* okruženju da nije radila (S_W^{CW}), a u *Load Flow* okruženju da je značaj rada ovog uređaja visok (S_P^{CW})
- visok značaj rada klima uređaja (S_P^{AC}).

Međusobne zavisnosti između rada uređaja ne mogu se u potpunosti objasniti na osnovu ovih analiza. Uloge karakteristika vezanih za mašinu za veš u donošenju odluka o radu klima uređaja nemaju jasno opravdanje i ne proizlazi iz očekivane logike energetskeg

upravljanja. Posebno je nelogično što značaj rada klima uređaja, koji bi u realnim uslovima trebalo da bude dominantan faktor, pokazuje ograničen uticaj na odluku o aktiviranju ovog uređaja.



Slika 54. DDQN model: uticaj karakteristika stanja na akciju “uključi klima uređaj” tokom letnjeg dana

Na slikama 55 i 57 prikazan je uticaj karakteristika stanja na akciju uključivanja mašine za sudove kod PPO i DDQN modela, tim redosledom.

Karakteristike koje utiču na odluku PPO modela da pokrene mašinu za sudove u oba simulaciona okruženja u letnjem periodu su:

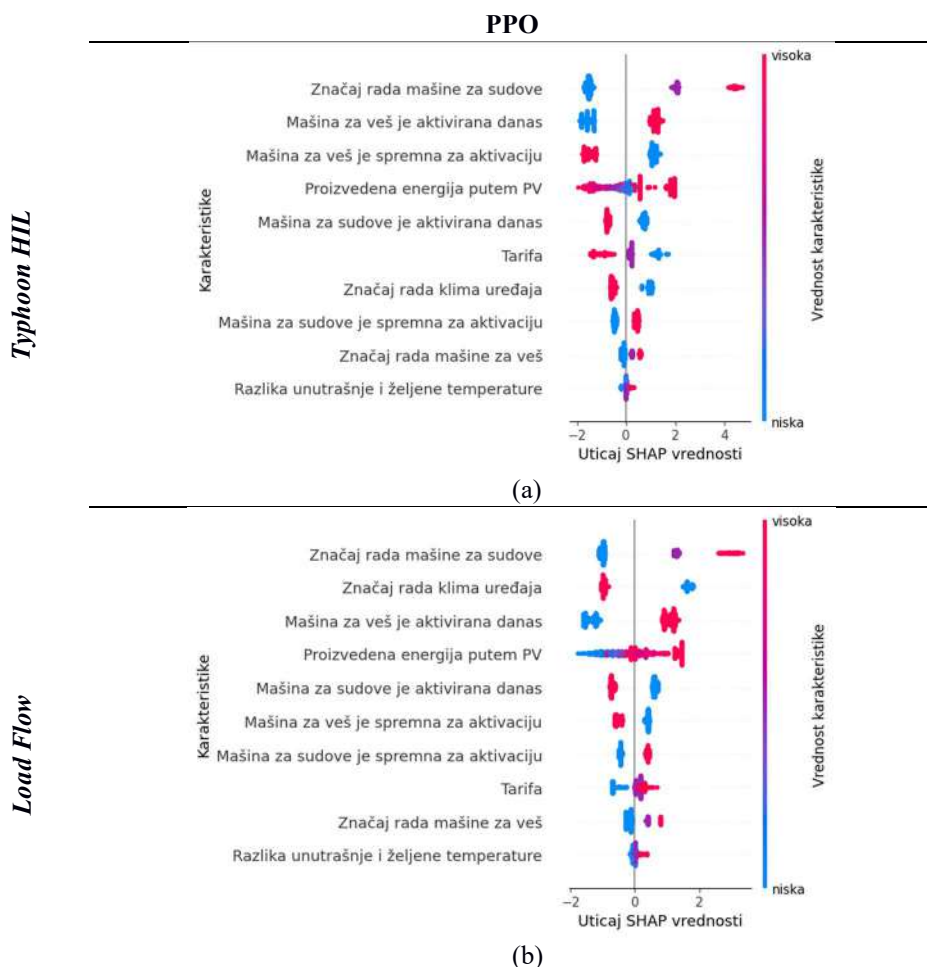
- visok značaj rada mašine za sudove (S_p^{DW}),
- niska vrednost za indikator da je veš mašina spremna za aktivaciju (S_L^{CW}),
- visoka vrednost za indikator da je veš mašina danas radila (S_W^{CW}),
- niska PV proizvodnja (S^{PV}).

Dodatno, manji uticaj ima spremnost mašine za sudove za rad, gde je visoka vrednost za indikator da je mašina za sudove spremna za rad (S_L^{DW}) i niska vrednost za indikator da mašina nije radila danas (S_L^{DW}). Niska vrednost karakteristike S_W^i , gde $i \in \{CW, DW\}$,

označava da mašina još uvek nije radila, dok visoka označava da jeste. Slično, niska vrednost karakteristike S_L^i označava da mašina nije spremna za rad, a visoka vrednosti da jeste.

Vrednosti ovih karakteristika se u najvećoj meri poklapaju sa stvarnim odlukama PPO modela o aktiviranju mašine za sudove u posmatranom letnjem danu (slika 26). Model je, u skladu sa očekivanim obrascem, donosio odluke o aktiviranju uređaja nakon završetka rada mašine za veš i u periodima visokog značaja rada mašine za sudove.

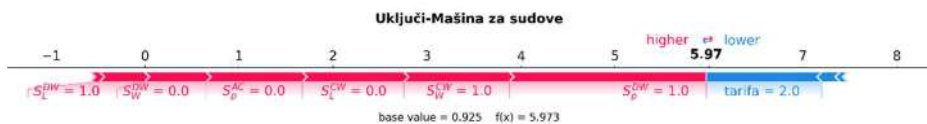
Ipak, kontraintuitivno je da niska PV proizvodnja ima pozitivan doprinos odluci o pokretanju mašine za sudove. Ovakvo ponašanje odstupa od očekivanog energetskog racionaliteta, budući da bi veća raspoloživa proizvodnja solarne energije trebalo da bude podsticaj za aktiviranje potrošača.



Slika 55. PPO model: uticaj karakteristika stanja na akciju “uključi mašinu za sudove” tokom letnjeg perioda.

Na slici 56 je prikazan SHAP dijagram koji objašnjava odluku PPO modela da aktivira mašinu za sudove u konkretnom vremenskom koraku. Sve karakteristike koje doprinose odluci imaju logičan smer uticaja i u skladu su sa očekivanim obrascima ponašanja modela. Srednji značaj rada mašine za sudove (S_p^{DW}), spremnost mašine za sudove za rad

(S_L^{DW}, S_W^{DW}) i završetak rada mašine za veš (S_L^{CW}, S_W^{CW}) doprinose povećanju verovatnoće aktiviranja mašine za sudove. Nasuprot tome, visoka tarifa (*tarifa*) ima manji negativan doprinos odluci, što je u skladu sa očekivanim uticajem troškova energije. Sporna niska PV proizvodnja nema značajan uticaj na odluku. Time se potvrđuje da je odluka PPO modela u ovom primeru u potpunosti u skladu sa logikom energetski efikasnog i racionalnog upravljanja uređajem.



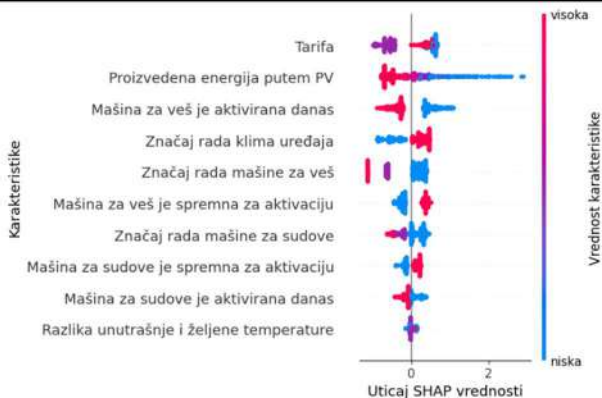
Slika 56. PPO model: Uticaj karakteristika na akciju “uključi mašinu za sudove” u *Typhoon HIL* simulacionom okruženju u letnjem danu za stanje u određenom vremenskom koraku.

Karakteristike koje utiču na odluku DDQN modela da aktivira mašinu za sudove u oba simulaciona okruženja su:

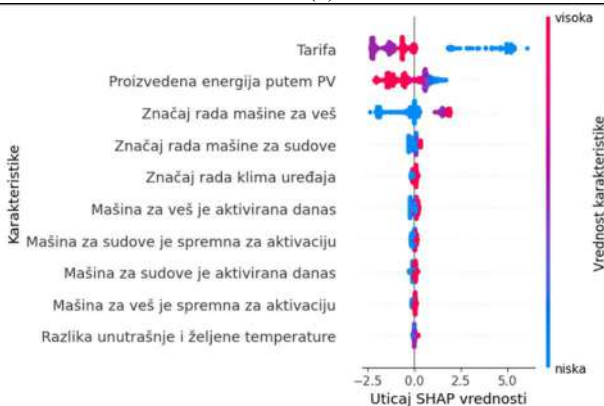
- niska tarifa,
- visoka PV proizvodnja električne energije (S_{PV}),
- indikator da mašina za veš nije radila danas (S_W^{CW}),
- visok značaj rada klima uređaja (S_P^{AC}) i
- nizak značaj rada mašine za veš u *Typhoon HIL* simulacionom okruženju i visok značaj rada mašine za veš u *Load Flow* simulacionom okruženju (S_P^{CW}).

Odluke DDQN modela prikazane na slici 26 uglavnom se poklapaju sa karakteristikama identifikovanim na globalnom nivou (slika 57) u oba simulaciona okruženja. Uređaji su se u oba okruženja aktivirali u periodima visoke proizvodnje električne energije. U *Typhoon HIL* okruženju aktivacija se javljala tokom niskog značaja rada klima uređaja, dok se u *Load Flow* okruženju dešavala u periodu niskog značaja rada mašine za veš.

Za razliku od globalne analize, u oba simulaciona okruženja model je mašinu za sudove aktivirao u periodu visoke, a ne niske tarife. U *Load Flow* okruženju dodatno odstupa i karakteristika značaja rada mašine za veš, jer se model nije ponašao u skladu sa globalnim obrascem prema kome se uređaj aktivira pri visokom značaju rada mašine za veš.



(a)



(b)

Slika 57. DDQN model: uticaj karakteristika stanja na akciju “uključi mašinu za sudove” tokom letnjeg perioda.

Karakteristike koje su imale pozitivan uticaj na odluku DDQN modela o aktiviranju mašine za sudove u *Typhoon HIL* simulacionom okruženju uključuju visoku tarifu (što je kontrainuitivno), visoku PV proizvodnju i spremnost za rad mašine za sudove. Negativan doprinos odluci imaju nizak značaj rada klima uređaja, činjenica da je mašina za sudove već radila, kao i (kontrainuitivno) srednji značaj rada mašine za sudove (slika 58).

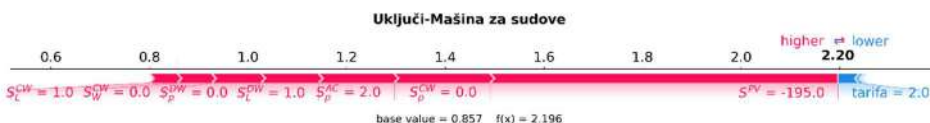
Iako je DDQN model u *Typhoon HIL* simulacionom okruženju uspeo da postigne relativno dobar balans između uštede energije i komfora korisnika, smer uticaja pojedinih faktora na odluku nije u skladu sa očekivanim obrascima energetske racionalnog ponašanja.



Slika 58. DDQN model: Uticaj karakteristika na akciju “uključi mašinu za sudove” u *Typhoon HIL* simulacionom okruženju u letnjem danu za stanje u određenom vremenskom koraku.

Na slici 59 prikazana je konkretna odluka DDQN modela u *Load Flow* simulacionom okruženju da aktivira mašinu za sudove. U ovom primeru, model je imao pretežno pozitivne faktore: visoku PV proizvodnju (S^{PV}), nizak značaj rada mašine za veš (S_P^{CW}), visok značaj rada klima uređaja (S_P^{AC}) i spremnost mašine za sudove za rad (S_L^{DW}). Takođe, logičan negativan uticaj je imala visoka tarifa (*tarifa*). Međutim, uočen je kontrainuitivan pozitivan doprinos niskog značaja rada mašine za sudove, što odstupa od očekivanog ponašanja sistema. Dodatno, karakteristika visoke vrednosti značaja rada mašine za veš, koja se na globalnom nivou izdvojila kao uticajna (slika 57 (b)), nije imala značajan uticaj na odluku modela u ovom slučaju.

Za razliku od PPO modela, DDQN u *Load Flow* simulacionom okruženju nije uspeo da uspostavi ravnotežu između uštede električne energije i očuvanja komfora stanara.



Slika 59. DDQN model: Uticaj karakteristika na akciju “uključi mašinu za sudove” u *Load Flow* simulacionom okruženju u letnjem danu za stanje u određenom vremenskom koraku.

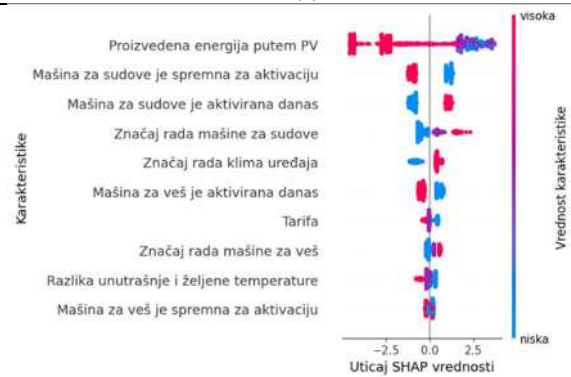
Karakteristike sa značajnim uticajem na odluku PPO modela da aktivira mašinu za veš u oba simulaciona okruženja su:

- visoka i niska PV proizvodnja (S^{PV}),
- u *Typhoon HIL* simulacionom okruženju, indikator spremnosti za rad mašine za veš (S_W^{CW} i S_L^{CW})
- u *Load Flow* simulacionom okruženju je manji uticaj imao i indikator da je mašina za sudove radila danas (S_W^{DW}).

Na globalnom nivou (slika 60), navedene karakteristike pokazuju dosledan uticaj na odluku PPO modela, dok se na nivou jednog odabranog letnjeg dana (slika 27) uočava njihovo poklapanje sa stvarnim upravljanjem uređaja. Jedino odstupanje predstavlja indikator da je mašina za sudove radila, koji nema značajan uticaj na odluku o aktiviranju mašine za veš.

Typhoon HIL

(a)

Load Flow

(b)

Slika 60. PPO model: uticaj karakteristika stanja na akciju “*uključi veš mašinu*” tokom letnjeg dana.

Na slici 61 prikazan je uticaj karakteristika na odluku PPO modela da aktivira mašinu za veš u *Typhoon HIL* simulacionom okruženju u letnjem danu. Pozitivni faktori bili su visoka PV proizvodnja (S^{PV}), indikator da mašina za veš nije dana radila (S_W^{CW}), da je mašina za veš spremna za rad (S_L^{CW}), visoka tarifa (*tarifa*) i nizak značaj rada mašine za sudove (S_P^{DW}). Negativan uticaj su imale karakteristike da mašina za sudove nije radila danas (S_W^{DW}), visok značaj rada klima uređaja (S_P^C) i nizak značaj rada mašine za veš (S_P^{CW}). Kontraintuitivni uvid da je pozitivan utican na aktiviranje mašine za veš ima karakteristika da je mašina za sudove radila (slika 60 (a)) se ne pojavljuje u predikciji.



Slika 61. PPO model: Uticaj karakteristika na akciju “*uključi mašinu za veš*” u *Typhoon HIL* simulacionom okruženju u letnjem danu za stanje u određenom vremenskom koraku.

Na slici 62 prikazane su karakteristike koje su uticale na odluku PPO modela da aktivira mašinu za veš u jednom određenom vremenskom koraku. Pozitivan doprinos odluci imale su visoka PV proizvodnja (S^{PV}), visok značaj rada klima uređaja (S_P^C) i činjenica

da mašina za veš još uvek nije bila aktivna (S_W^{CW}). Nasuprot tome, negativan uticaj na odluku imale su karakteristike koje ukazuju da je mašina za sudove spremna za rad i nije radila (S_L^{DW} i S_W^{DW}), visoka tarifa (*tarifa*) električne energije i nizak značaj rada mašine za sudove (S_p^{DW}). Sporni uvid iz globalne analize da pozitivan uticaj na odluku uključivanja mašine za veš ima činjenica da je mašina za sudove radila (slika 60 (b)) nije imala u ovom primeru uticaja na odluku o aktiviranju mašine za veš.

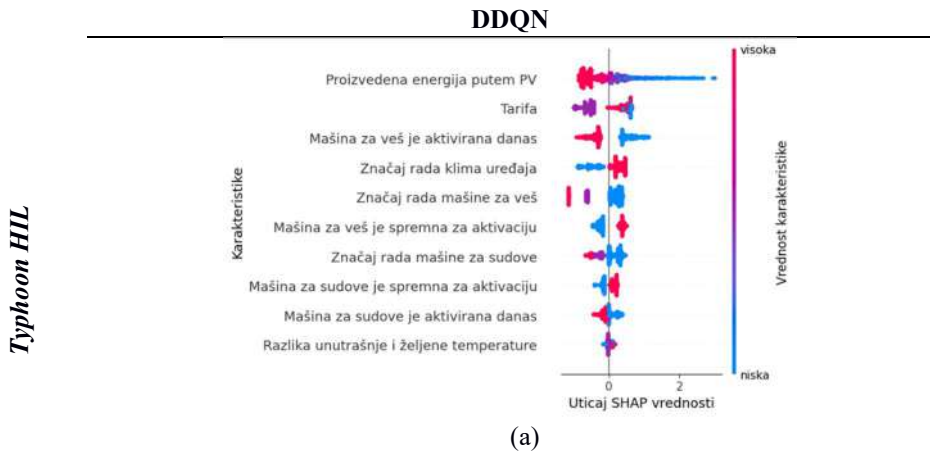


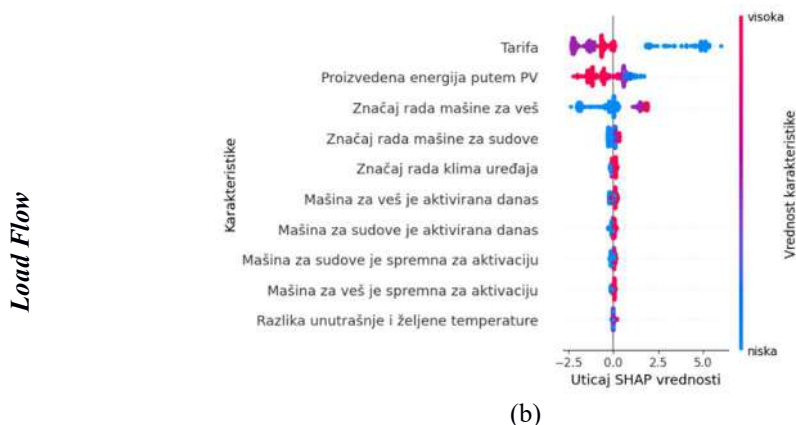
Slika 62. PPO model: Uticaj karakteristika na akciju “uključi mašinu za veš” u *Load Flow* simulacionom okruženju u letnjem danu za stanje u određenom vremenskom koraku.

Karakteristike sa značajnim uticajem na odluku DDQN modela da aktivira mašinu za veš u oba simulaciona okruženja su:

- visoka PV proizvodnja električne energije (S^{PV}),
- niska i visoka vrednost tarife u *Typhoon HIL* simulacionom okruženju i isključivo niske vrednosti tarife u *Load Flow* simulacionom okruženju (*tarifa*),
- indikator da mašina za veš nije radila danas u *Typhoon HIL* simulacionom okruženju (S_W^{CW}) i
- visok značaj rada mašine za veš u *Load Flow* simulacionom okruženju (S_p^{CW}).

Sve karakteristike identifikovane na globalnom nivou (slika 63), osim visokog značaja rada mašine za veš u *Load Flow* simulacionom okruženju, se poklapaju sa uočenom logikom upravljanja mašine za veš od strane obučenog DDQN modela (slika 27).





Slika 63. DDQN model: uticaj karakteristika stanja na akciju “uključiti mašinu za veš” tokom letnjeg perioda.

Na slici 64 prikazano je aktiviranje mašine za veš od strane obučenog DDQN modela u jednom određenom vremenskom koraku. Pozitivan doprinos odluci imale su karakteristike visoke PV proizvodnje (S^{PV}), niskog značaja rada mašine za sudove (S_p^{DW}), spremnosti za rad mašine za veš (S_L^{DW} i S_W^{DW}) i (kontraintuitivno) nizak značaj rada mašine za veš (S_p^{CW}). Negativan uticaj imale su visoka tarifa (*tarifa*) i nizak značaj rada klima uređaja (S_p^{AC}).

Sporna karakteristika visokog značaja rada mašine za veš, koja je na globalnom nivou imala pozitivan doprinos odluci (slika 63 (b)), u ovom slučaju nije imala značajan uticaj. DDQN model je u ovom primeru pokazao da pridaje veći značaj uštedi električne energije na račun komfora stanara.



Slika 64. DDQN model: Uticaj karakteristika na akciju “uključiti mašinu za veš” u Load Flow simulacionom okruženju u letnjem danu za stanje u određenom vremenskom koraku.

6.1.4.1.2.1 Zaključak na nivou odluka u letnjem periodu

Analiza rezultata pokazuje da PPO model u oba simulaciona okruženja donosi odluke koje su u većoj meri usklađene sa logikom energetski efikasnog i komforanog upravljanja uređajima. Model je pokazao doslednost u ponašanju gde su se odluke o aktiviranju uređaja najčešće poklapale sa visokim značajem rada pojedinačnih uređaja i završetkom rada drugih potrošača, uz efikasno korišćenje dostupne solarne energije. Iako su povremeno uočeni kontraintuitivni doprinosi niske PV proizvodnje, njihov uticaj nije značajno narušio racionalnost odluka. PPO je zadržao stabilnost i u letnjem periodu, posebno u *Typhoon HIL* okruženju, gde su odluke bile konzistentnije i bolje prilagođene dinamičkim uslovima proizvodnje i potrošnje.

Suprotno tome, DDQN model je u oba okruženja pokazao veći stepen varijabilnosti i manju sposobnost uspostavljanja balansa između uštede energije i komfora korisnika. U

Typhoon HIL okruženju model je povremeno postizao zadovoljavajuće rezultate, ali sa nelogičnim smerom uticaja pojedinih karakteristika, poput visoke tarife i niskog značaja rada uređaja, koje su pozitivno doprinosile odluci. U *Load Flow* okruženju odstupanja su bila izraženija. Model je davao prioritet uštedi električne energije u odnosu na komfor, što je dovelo do odluka koje nisu u skladu sa očekivanim obrascima ponašanja sistema pametne kuće.

Na osnovu sprovedene analize može se zaključiti da je PPO model ostvario bolje rezultate od DDQN modela u pogledu stabilnosti, racionalnosti i usklađenosti odluka sa ciljevima optimizacije sistema. PPO je pokazao veću sposobnost generalizacije između simulacionih okruženja, dok je DDQN pokazao ograničenja u adaptaciji i sklonost nelogičnim obrascima odlučivanja. Time se potvrđuje da PPO model predstavlja pogodniji pristup za upravljanje pametnim kućnim sistemima tokom letnjeg perioda, gde je potrebno istovremeno održati energetske efikasnost i komfor korisnika.

6.1.4.1.3 Zaključak interpretacije odluka na nivou različitih godišnjih doba

Rezultati sprovedenih eksperimenata pokazuju da je PPO model u oba sezonska perioda ostvario stabilnije i racionalnije upravljanje uređajima u odnosu na DDQN model. U zimskom periodu PPO je pokazao sposobnost višeciljne optimizacije, balansirajući između uštede energije i održavanja komfora stanara. Model je dosledno reagovao na ključne faktore kao što su značaj rada uređaja, termalni komfor i raspoloživa PV proizvodnja, pri čemu su odluke bile u skladu sa logikom realnog energetskeg sistema. U *Typhoon HIL* i *Load Flow* okruženju PPO je pokazao slične obrasce ponašanja, što ukazuje na dobru generalizaciju modela i otpornost na promenu uslova okruženja.

U letnjem periodu PPO model je zadržao sličan obrazac odlučivanja. Odluke su se uglavnom poklapale sa periodima visokog značaja rada uređaja i završetkom rada drugih potrošača, uz efikasno korišćenje solarne energije. Povremeno su se javljali kontrainuitivni doprinosi niske PV proizvodnje, ali bez značajnog uticaja na performanse. U *Typhoon HIL* okruženju PPO je pokazao najveću stabilnost i konzistentnost, dok su u *Load Flow* okruženju uočene manje varijacije u uticaju pojedinačnih karakteristika, ali bez narušavanja energetske racionalnosti odluka.

Nasuprot tome, DDQN model pokazao je veću osetljivost na pojedinačne karakteristike i manju sposobnost uspostavljanja balansa između troškova i komfora u oba perioda. U zimskom periodu tarifa je imala dominantan uticaj na odluke. U letnjem periodu model je pokazao veći stepen varijabilnosti i nelogične doprinose karakteristika poput visoke tarife i niskog značaja rada uređaja. U *Typhoon HIL* okruženju DDQN je povremeno ostvarivao zadovoljavajuće odluke, dok je u *Load Flow* okruženju pokazao slabiju adaptaciju i odstupanja od očekivanih obrazaca ponašanja.

Na osnovu ukupne analize može se zaključiti da PPO model daje bolje rezultate u odnosu na DDQN u pogledu stabilnosti, interpretabilnosti i sposobnosti generalizacije između različitih sezonskih uslova i simulacionih okruženja. PPO je pokazao uravnotežen pristup između energetske efikasnosti i komfora, dok DDQN model pokazuje manju sposobnost integracije više ciljeva u procesu odlučivanja, što ga čini manje pogodnim za kompleksne zadatke upravljanja pametnim kućnim sistemima.

6.2 Analiza rezultata velikog jezičkog modela

U ovom potpoglavlju prikazani su rezultati upravljanja uređajima u pametnoj kući ostvareni pomoću pretreniranih velikih jezičkih modela. U potpoglavlju 6.2.1 prikazane su vrednosti funkcije nagrade koju veliki jezički modeli ostvaruju na 23. danu u januaru i junu, istom danu na kom su u potpoglavlju 6.1.2 analizirani modeli učenja potkrepljivanjem. Za svako simulaciono okruženje prikazana je posebna tabela, pri čemu su u svakoj tabeli objedinjeni rezultati za zimski i letnji dan. Potpoglavlje 6.2.2 analizira odluke modela u tim danima, gde su posebno izdvojene odluke modela vezane za kontrolisanje unutrašnje temperature i upravljanje radom mašine za pranje sudova i mašine za veš.

Zbog ograničenja resursa, odluke donesene od strane RL LLM modela su analizirane samo u *Load Flow* simulacionom okruženju.

6.2.1 Analiza vrednosti funkcije nagrade na nivou simulacionog okruženja

U ovom potpoglavlju analizirane su vrednosti funkcije nagrade koje veliki jezički modeli ostvaruju na 23. danu u januaru i junu, koji neposredno slede periodu treniranja modela. Zasebno su analizirane i komponente funkcije nagrade – udobnost i cena električne energije. Cilj ove analize je da se ispita mogućnost primene pretreniranih velikih jezičkih modela za upravljanje uređajima u pametnoj kući.

U tabelama 18 i 19 rezultati su prikazani na nivou simulacionog okruženja (*Load Flow* i *Typhoon HIL*) i obuhvataju ostvarene rezultate u zimskom i letnjem danu. Rezultati se ne mogu porediti između različitih simulacionih okruženja jer konfiguracije uređaja nije identična na oba simulatora.

U januaru (tabela 18), u *Load Flow* simulacionom okruženju, KBS LLM je postigao bolje rezultate u poređenju sa RL LLM u pogledu svih posmatranih metrika. U januarskom danu, KBS LLM je ostvario ukupnu funkciju nagrade veću za 0,24 €, što odgovara unapređenju od oko 13%.

U junu je KBS LLM ostvario bolje rezultate u pogledu cene električne energije i ukupne nagrade, dok je RL LLM imao bolji rezultat u pogledu udobnosti. Ipak, KBS LLM ostvario je bolji balans između udobnosti i uštede električne energije, odnosno, ostvario je višu ukupnu nagradu. U letnjem danu, KBS LLM ostvaruje ukupnu nagradu veću za 0,05 €, tj. oko 5% unapređenja.

Tabela 18. *Load Flow* okruženje: vrednost funkcije nagrade ostvarene na testnim danima (23. januar i jun). Negativne vrednosti označavaju da je potrošnja energije bila veća od proizvodnje, dok pozitivne vrednosti označavaju da je proizvodnja električne energije veća od potrošnje.

	Model	Udobnost (27) sa kaznama za nepoštovanje želja stanara [€]	Cena električne energije (32) [€]	Zbir komponenti udobnosti i cene [€]
Januar	KBS	6,26	-4,16	2,10
	LLM			
	RL	6,05	-4,19	1,86
	LLM			

<i>Jun</i>	KBS LLM	1,14	-0,25	0,89
	RL LLM	2,18	-1,34	0,84

Tabela 19 prikazuje vrednosti KBS LLM u *Typhoon HIL* simulacionom okruženju. Zbog nedostatka resursa, RL LLM model nije pokrenut u datom okruženju.

Tabela 19. *Typhoon HIL* okruženje: vrednost funkcije nagrade ostvarene na testnim danima (23. januaru i junu). Negativne vrednosti označavaju da je potrošnja energije bila veća od proizvodnje, dok pozitivne vrednosti označavaju da je proizvodnja električne energije veća od potrošnje.

	Model	Udobnost (27) sa kaznama za nepoštovanje želja stanara [€]	Cena električne energije (32) [€]	Zbir komponenti udobnosti i cene [€]
<i>Januar</i>	KBS LLM	5,41	-3,53	1,88
<i>Jun</i>	KBS LLM	1,04	0,60	1,63

Rezultati pokazuju da je KBS LLM ostvario više vrednosti funkcije nagrade u poređenju sa RL LLM modelom, uz unapređenja od oko 13% u zimskom i 5% u letnjem danu. KBS LLM je postigao bolji balans između komfora i troškova električne energije, čime je obezbeđena efikasnija upotreba resursa. Dodatno, KBS LLM pristup zahteva znatno manje resursa, jer ne zahteva višestruko obučavanje niti kontinuirano slanje upita velikom jezičkom modelu, što ga čini pogodnijim za primenu u realnim pametnim okruženjima.

6.2.2 Analiza odluka modela na nivou dana

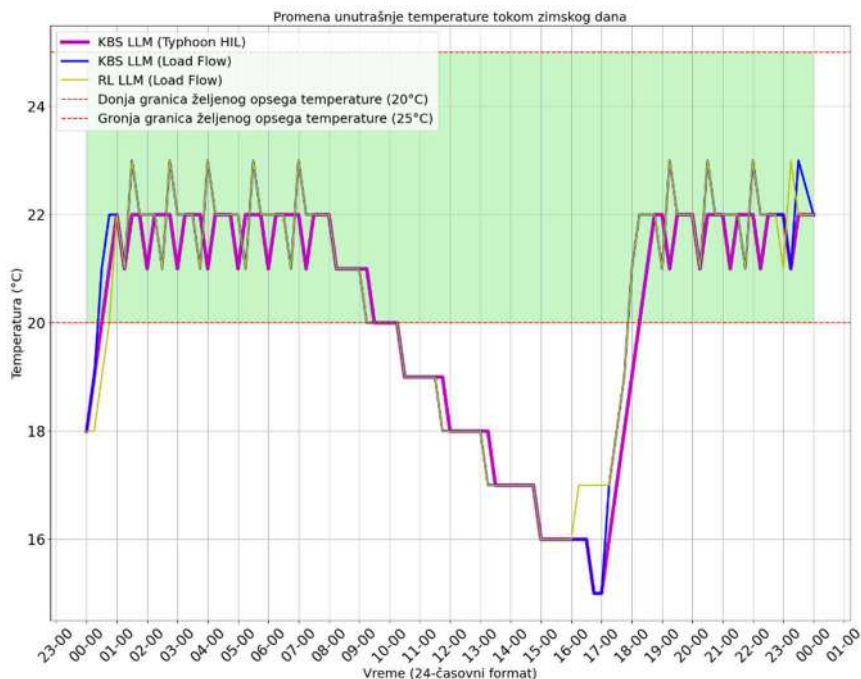
Analiza odluka modela treba da ukaže mogućnost primene velikog jezičkog modela u upravljanju uređajima pametne kuće sa ciljem optimizacije potrošnje električne energije uz istovremeno održavanje udobnosti stanara. Analiza je grupisana po sezonama: zimski dan (potpoglavlje 6.2.2.1) i letnji dan (potpoglavlje 6.2.2.2), pri čemu se zasebno posmatra kontrola klima uređaja, rad mašine za pranje sudova i rad mašine za veš. Posmatrani su isti 23. dan u januaru i junu kao i u analizi modela zasnovani na učenju potkrepljivanjem.

Zbog nedostatka resursa u analizi neće se prikazati rezultati RL LLM *Typhoon HIL* simulacionom okruženju.

6.2.2.1 Analiza rada uređaja u zimskom danu

Slika 65 prikazuje kako modeli kontrolišu unutrašnju temperaturu tokom zimskog dana. Bez obzira na intenzitet potrošnje klima uređaja u različitim simulacionim okruženjima, modeli na sličan način upravljaju klima uređajem. RL LLM se razlikuje od ostalih modela po tome što aktivira klima uređaj neposredno pre početka opsega velikog značaja temperature (u 16:00), što je u skladu sa željama stanara iz tabele 13. Ostali modeli

aktiviraju klimu uređaj tačno na početku opsega velikog značaja temperature (u 17:00). Razlog zašto su oba modela u *Load Flow* simulacionom okruženju dostigla željeni temperaturni opseg u isto vreme (oko 18:00) je što je KBS LLM pokrenuo klimu uređaj 15 minuta ranije od RL LLM modela.



Slika 65. Kontrolisanje unutrašnje temperature u toku zimskog dana.

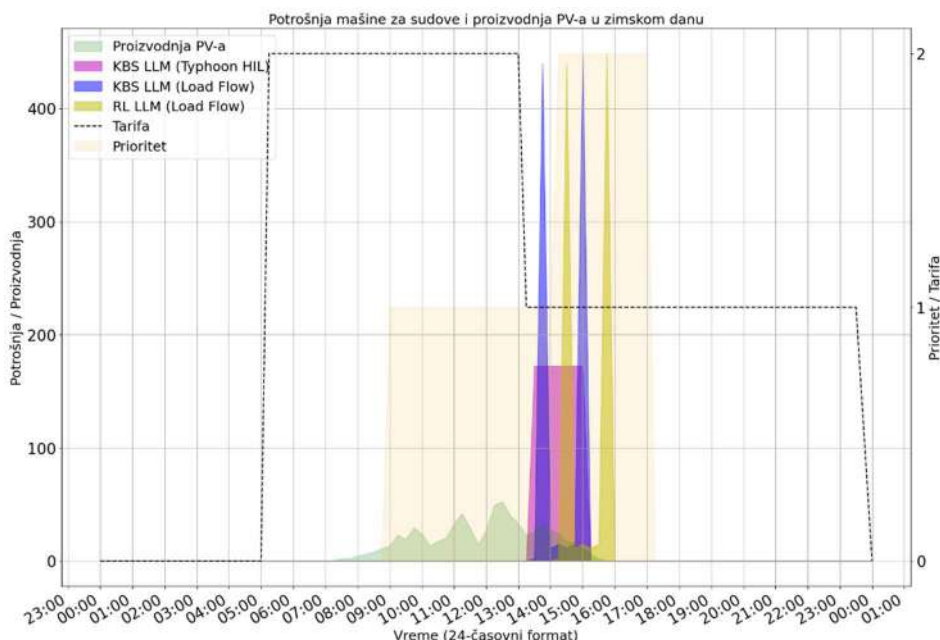
Na slici 66 uočava se da je KBS LLM postigao veću kumulativnu udobnost u odnosu na RL LLM jer je oko 13:00 vidljiv skok vrednosti kod KBS LLM modela, koji odgovara trenutku aktiviranja mašine za sudove (slika 67). Zbog toga RL LLM nije uspeo da dostigne ukupnu vrednost udobnosti koju je ostvario KBS LLM do kraja dana.



Slika 66. Udobnost u zimskom danu.

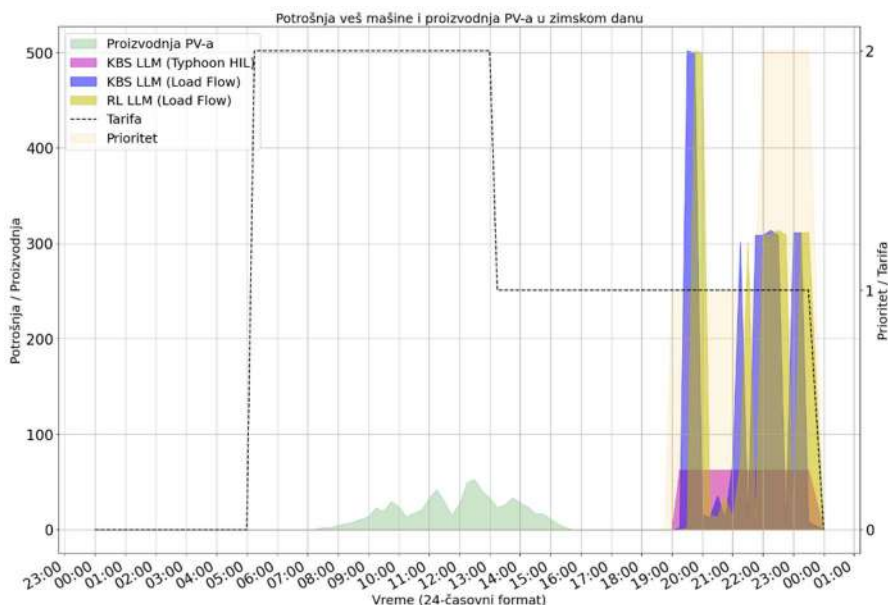
Slike 67 i 68 prikazuju način upravljanja mašinom za pranje sudova u zimskom danu. KBS LLM u oba simulaciona okruženja pokreće uređaj u periodu srednje tarife i visokog

značaja rada. RL LLM pokreće uređaj nešto kasnije kada ima manje proizvedene električne energije. Kao posledica, KBS LLM model postiže bolju uštedu električne energije u odnosu na RL LLM model, bez kompromitovanja udobnosti stanara (tabela 18). Takođe je rezultovalo u većoj vrednosti udobnosti od strane KBS LLM zbog definisane značaja rada u tabeli 14.



Slika 67. Upravljanje mašinom za sudove u zimskom danu.

Slika 68 ilustruje kako modeli upravljaju radom mašine za veš u zimskom danu. Svi modeli, nezavisno od simulacionog okruženja, pridodaju veći prioritet značaju rada uređaja u odnosu na uštedu električne energije. Ovo je posledica toga što u zimskom danu nema perioda visoke proizvodnje električne energije od strane solarnog panela.

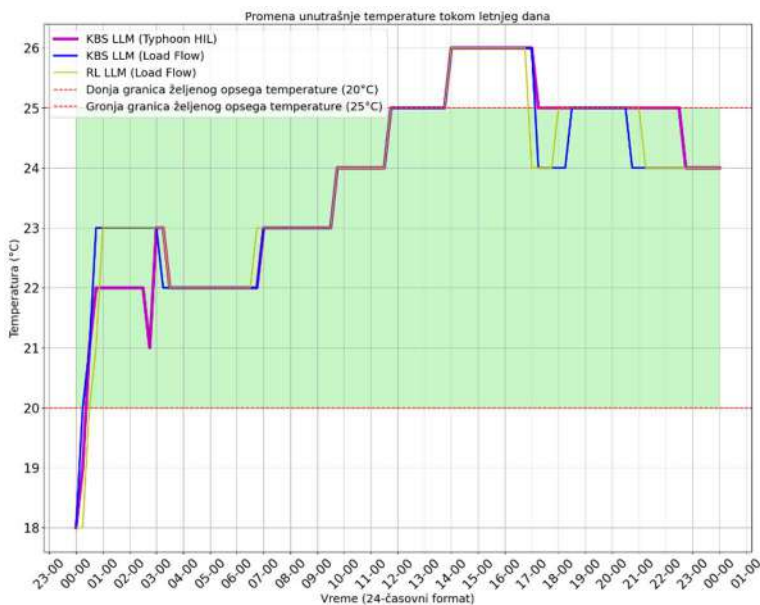


Slika 68. Upravljanje mašinom za veš u zimskom danu.

Prethodne slike i vrednosti metrika iz tabele 18 potvrđuju da oba modela imaju slične obrasce kontrole uređaja u *Load Flow* simulacionom okruženju.

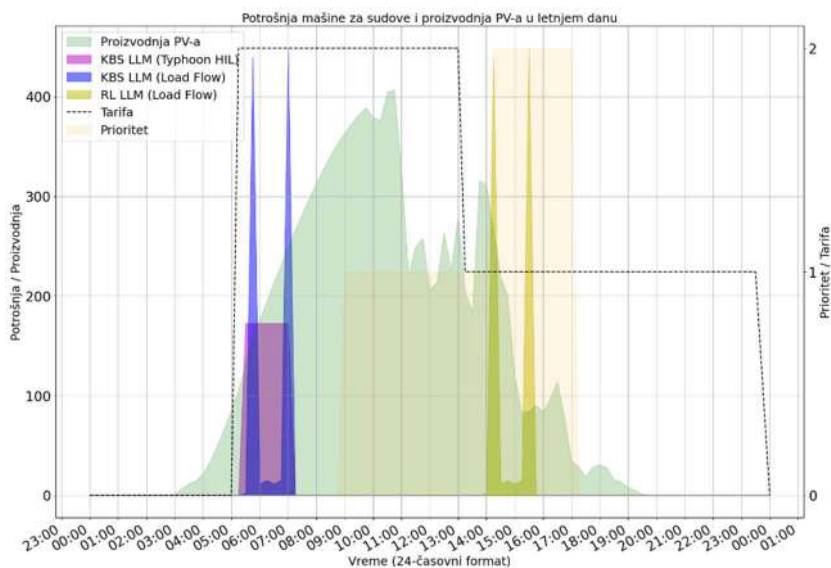
6.2.2.2 Analiza rada uređaja u letnjem danu

Slika 69 prikazuje kako modeli kontrolišu unutrašnju temperaturu tokom analiziranog letnjeg dana. Bez obzira na intenzitet potrošnje, modeli kontrolišu klima uređaj na gotovo identičan način. Modeli održavaju unutrašnju temperaturu u željenom opsegu do 8:00 nakon čega temperatura raste i modeli aktiviraju klimu uređaj tek oko 17:00, što je u skladu sa željama stanara iz tabele 13. RL LLM model, slično kao i u analiziranom zimskom danu, aktivira uređaj nešto ranije od KBS LLM modela, te brže dostigne željeni temperaturni opseg. Zbog ovoga je RL LLM ostvario bolje performanse u pogledu udobnosti od KBS LLM modela u *Load Flow* simulacionom okruženju (tabela 18).



Slika 69. Kontrolisanje unutrašnje temperature u toku letnjeg dana.

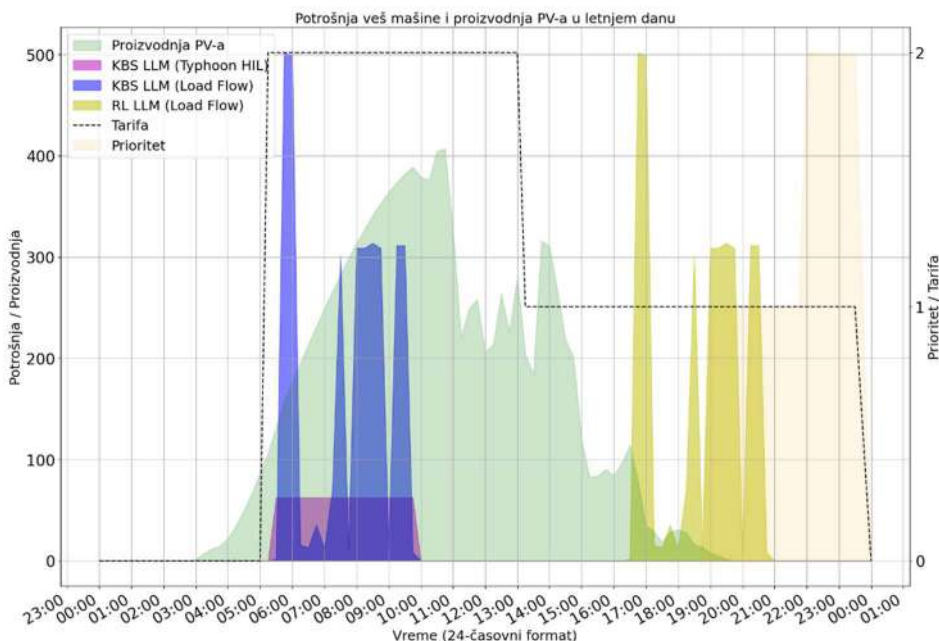
Slike 70 i 71 prikazuju način na koji modeli upravljaju mašinom za pranje sudova i mašinom za veš u analiziranom letnjem danu. Primećuje se da KBS LLM, nezavisno od simulacionog okruženja (*Typhoon HIL* i *Load Flow*), odnosno tipa potrošnje električne energije (konstantna ili promenljiva u jedinici vremena), inicira rad uređaja u istim vremenskim tačkama, dok se RL LLM više fokusira na udobnost stanara, što se odražava na konačne rezultate prikazane u tabeli 18.



Slika 70. Upravljanje mašinom za sudove u analiziranom letnjem danu.

Slika 71 ilustruje kako modeli upravljaju radom mašine za veš u analiziranom letnjem danu. Obrazac upravljanja KBS LLM u oba simulaciona okruženja identičan je kao kod

mašine za sudove u letnjem danu gde se uređaj pokreću u istom vremenskom intervalu koji se odlikuju sa viskom tarifom i istom proizvodnjom električne energije. Nasuprot tome, RL LLM uključuje uređaj nešto pre perioda srednjeg značaja rada, kako bi iskoristio povećanu proizvodnju električne energije od strane PV panela.



Slika 71. Upravljanje mašinom za veš u letnjem danu.

Rezultati pokazuju da je RL LLM ostvario bolje rezultate isključivo u pogledu udobnosti stanara. Kao što je prikazano na slikama (slika 69, slika 70 i slika 71), ovaj model pokreće uređaje u periodima srednjeg i visokog značaja rada, čime doprinosi višem nivou udobnosti stanara. Međutim, KBS LLM ostvario je bolje rezultate u pogledu uštede električne energije i ukupne nagrade jer aktivira uređaje u periodima visoke proizvodnje električne energije, iako po cenu manjeg nivoa udobnosti (slika 70 i slika 71).

6.2.2.3 Zaključak analize odluka model na nivou zimskog i letnjeg dana

Analiza upravljanja svih uređaja u zimskom danu je slična u oba simulaciona okruženja. Tokom zimskog dana, svi modeli se fokusiraju na udobnost stanara, i pokreću uređaje kada je značaj rada srednji ili visok.

Suprotno tome, u letnjem danu u kome je proizvodnja električne energije putem solarnog panela veća, KBS LLM pokreće uređaje izvan opsega srednjeg i visokog značaja rada, nezavisno od korišćenog simulacionog okruženja. Nasuprot tome, RL LLM, donosi odluke nezavisno od godišnjeg doba i proizvodnje solarnog panela, te uređaje aktivira isključivo kada postoji visok značaj rada.

6.3 Poređenje modela na nivou jednog dana sa KBS definisanim od strane ljudskog eksperta

U ovom potpoglavlju prikazane su tabele koje sumiraju sve prethodno prikazane rezultate u pogledu udobnosti, potrošnje električne energije i ukupne nagrade za sve uređaje. Svi

modeli su evaluirani na istim danima, odnosno na zimskom i letnjem testnom danu koji slede neposredno nakon perioda treniranja. Kao referentni model korišćen je KBS definisan od strane ljudskog eksperta.

Za zimski dan u *Load Flow* simulacionom okruženju, KBS definisan od strane ljudskog stručnjaka i KBS LLM su postigli najbolje rezultate (tabela 20). Sledeći najuspešniji model je RL LLM, sa metrikom udobnosti nižom za 0,21 € i za 0,03 € većom uštedom električne energije, što čini da je ukupna nagrada gora za 0,24 € (umanjenje od oko 13%). DDQN model ostvaruje veću uštedu električna energije jer nije aktivirao mašinu za već, zbog čega je i metrika udobnosti niža. Slični rezultati uočeni su i u *Typhoon HIL* simulacionom okruženju (tabela 21).

Za odabrani letnji dan, referentni model ostvario je najbolje rezultate u pogledu udobnosti u oba simulaciona okruženja (tabela 20 i tabela 21). U *Load Flow* simulacionom okruženju, PPO model imao je najbolje performanse ukupne nagrade za 0,61 € više u odnosu na KBS definisanog od strane ljudskog eksperta. U *Typhoon HIL* simulacionom okruženju, DDQN model je pokazao najbolje rezultate u pogledu cene električne energije za 2,32 € i ukupne nagrade za 1,72 € u odnosu na KBS definisanog od strane ljudskog eksperta.

Tabela 20. Metrike performansi za zimski i letnji dan u *Load Flow* simulacionom okruženju, ukazujući na potrošnju električne energije, udobnost i ukupnu nagradu za sve uređaje zajedno, izražene u evrima. Negativne vrednosti označavaju da je potrošnja energije bila veća od proizvodnje, dok pozitivne vrednosti označavaju da je proizvodnja električne energije veća od potrošnje.

	Model	Udobnost (27) sa kaznama za nepoštovanje želja stanara [€]	Cena električne energije (32) [€]	Zbir komponenti udobnosti i cene [€]
Januar	DDQN	0,49	-3,11	-2,62
	PPO	5,80	-4,20	1,60
	KBS LLM	6,26	-4,16	2,10
	RL LLM	6,05	-4,19	1,86
	KBS definisan od strane ljudskog eksperta	6,26	-4,16	2,10
Jun	DDQN	1,09	0,72	1,81
	PPO	1,39	0,67	2,10
	KBS LLM	1,14	-0,25	0,89
	RL LLM	2,18	-1,34	0,84
	KBS definisan od strane ljudskog eksperta	2,84	-1,35	1,49

Tabela 21. Metrike performansi za zimski i letnji dan u *Typhoon HIL* simulacionom okruženju, ukazujući na potrošnju električne energije, udobnost i ukupnu nagradu za sve uređaje zajedno, izražene u evrima. Negativne vrednosti označavaju da je potrošnja energije bila veća od proizvodnje, dok pozitivne vrednosti označavaju da je proizvodnja električne energije veća od potrošnje.

	Model	Udobnost (27) sa kaznama za nepoštovanje želja stanara [€]	Cena električne energije (32) [€]	Zbir komponenti udobnosti i cene [€]
Januar	DDQN	5,30	-3,63	1,67
	PPO	5,22	-3,61	1,61
	KBS LLM	5,41	-3,53	1,88
	KBS definisan od strane ljudskog eksperta	5,41	-3,53	1,88
Jun	DDQN	1,36	1,62	2,98
	PPO	1,28	1,43	2,71
	KBS LLM	1,04	0,60	1,63
	KBS definisan od strane ljudskog eksperta	1,96	-0,70	1,26

KBS definisan od strane stručnjaka je više orijentisan ka maksimizaciji udobnosti stanara, što ističe potrebu za automatizacijom sistema upravljanja uređajima u pametnim kućama u kontekstu više-ciljne optimizacije.

PPO modeli su imali konzistentno dobre rezultate bliske najboljem rezultatu, gde su rezultati PPO modela u poređenju sa najboljim rezultatima niži u zimskom periodu za 23% u *Load Flow* i za 14% u *Typhoon HIL* simulacionim okruženjima.

U letnjem periodu u *Load Flow* simulacionom okruženju PPO model je ostvario najbolje rezultate.

Takođe, PPO modeli su pokazali da dobro funkcionišu u situacijama gde je neophodno detaljnije istražiti okruženje i razviti kompleksnije strategije, što ih čini pogodnijim za složenije i dinamičnije scenarije (Fuente and Guerra 2024). U letnjem periodu, RL modeli u oba simulaciona okruženja su više orijentisana ka uštedi električne energije i aktiviraju mašinu za sudove i veš u periodu više proizvodnje električne energije što je uticalo na niže vrednosti udobnosti.

6.4 Ispitivanje robusnosti modela

Prethodne analize su sprovedene na podacima koji se odnose na jedan dan. Međutim, izdvojeni dan može posedovati specifične karakteristike koje ga čine atipičnim, odnosno može biti “povoljan” ili “nepovoljan” za primenu modela. Kako bi se umanjio rizik donošenja zaključaka zasnovanih na specijalnom slučaju i obezbedila pouzdanija procena performansi, sprovedena je evaluacija robusnosti modela njihovom primenom

na dužem vremenskom periodu, koji obuhvata sedam uzastopnih dana neposredno nakon faze obučavanja modela. U zimskom periodu analizirani su dani koji slede nakon 23. januara, dok su u letnjem periodu analizirani dani koji slede nakon 23. juna.

U analizi se zasebno razmatraju tri metrike: udobnost, cena električne energije i ukupna nagrada. U tabelama 22 – 25 prikazane su srednje vrednosti (eng. *mean*) i standardne devijacije (eng. *standard deviation*, *std*) ovih metrika. Istaknute vrednosti označavaju najbolje performanse u okviru pojedinačne metrike, pri čemu veća srednja vrednost i manja standardna devijacija ukazuju na bolje performanse modela i veću robusnost u procesu donošenja odluka.

Tabela 22 prikazuje rezultate koje modeli postižu u *Load Flow* simulacionom okruženju u januaru, dok slika 72 daje njihov grafički prikaz. KBS definisan od strane ljudskog eksperta ostvario je najstabilnije i najuspešnije rezultate u pogledu skoro svih razmatranih metrika, isključujući cenu električne energije. U pogledu ukupne nagrade, PPO i KBS LLM modeli postižu rezultate konkurentne rezultatima KBS modela definisanog od strane ljudskog eksperta, ostvarujući ukupnu nagradu manju za 23% i 15%. U metrici udobnosti beleže približne performanse kao referentni model (udobnost je umanjena za 10% i 6%). Najslabije performanse pokazao je DDQN model, koji ostvaruje 178% manju ukupnu nagradu i 93% manju udobnost u poređenju sa ljudski definisanim KBS modelom. U pogledu električne energije, iako DDQN ostvaruje najbolje rezultate, oni se ne mogu se uporediti sa rezultatima ostalih modela jer je DDQN model u svega 50% slučajeva uspešno aktivirao mašinu za sudove i veš u toku dana. Izuzimajući DDQN model, je PPO model je najuspešniji po pitanju uštede električne energije, nakon čega slede KBS LLM i KBS definisan od strane ljudskog eksperta, koji ostvaruju 2% i 4% manju uštedu električne energije u poređenju sa rezultatima PPO modela.

Tabela 22. *Load Flow* okruženje: srednje vrednosti (*mean*) i standardne devijacije (*std*) metrika udobnosti, cene električne energije i ukupne nagrade, koje modeli ostvaruju tokom sedam testnih dana u januaru. Negativne vrednosti označavaju da je potrošnja energije bila veća od proizvodnje, dok pozitivne vrednosti označavaju da je proizvodnja električne energije veća od potrošnje.

Model/ metrika	Udobnost (mean ± std)	Cena električne energije (mean ± std)	Ukupna nagrada (mean ± std)	Broj pokretanja uređaja (od 14)
DDQN	0,46 ± 0,44	-3,11 ± 0,32	-2,64 ± 0,13	7
PPO	5,70 ± 0,45	-4,10 ± 0,34	1,60 ± 0,14	14
KBS LLM	5,94 ± 0,91	-4,18 ± 0,40	1,77 ± 0,55	14
KBS definisan od strane ljudskog eksperta	6,32 ± 0,55	-4,25 ± 0,34	2,08 ± 0,24	14

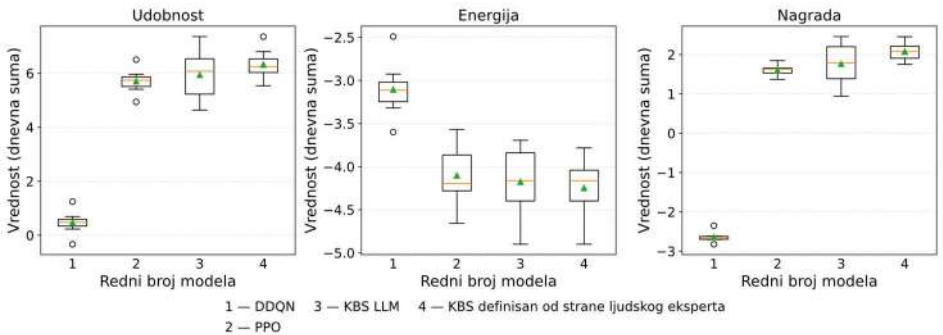
Tabela 23 prikazuje rezultate koje modeli postižu u *Typhoon HIL* simulacionom okruženju u januaru, dok slika 73 daje njihov grafički prikaz. KBS definisan od strane ljudskog eksperta ostvario je najstabilnije i najuspešnije rezultate u pogledu svih razmatranih metrika osim cene električne energije. U pogledu ukupne nagrade, PPO i KBS LLM modeli postižu rezultate konkurentne rezultate KBS modela definisanog od strane ljudskog eksperta, ostvarujući ukupnu nagradu manju za 15% i 3%. U metrici udobnosti PPO i KBS LLM modeli ostvaruju umanjene vrednosti za 5% i 3%. Najslabije

performanse pokazao je DDQN model, koji ostvaruje 132% manju ukupnu nagradu i 47% manju udobnost u poređenju sa ljudski definisanim KBS modelom. U pogledu električne energije, iako je DDQN ponovo uštedeo energiju time što nije uspešno aktivirao uređaje, najbolji rezultat je ostvario KBS LLM. Nakon tog modela, po pitanju uštede električne energije slede KBS definisan od strane ljudskog eksperta i PPO sa vrednostima umanjenim za 2,5% i 0,8%.

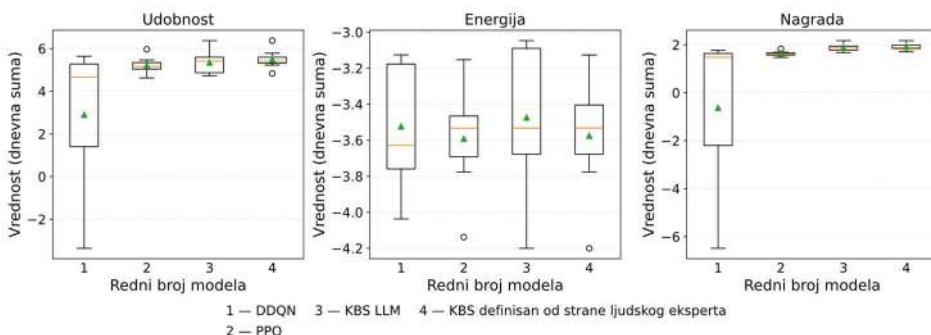
Tabela 23. *Typhoon HIL* okruženje: srednje vrednosti (*mean*) i standardne devijacije (*std*) metrika udobnosti, cene električne energije i ukupne nagrade, koje modeli ostvaruju tokom sedam testnih dana u januaru. Negativne vrednosti označavaju da je potrošnja energije bila veća od proizvodnje, dok pozitivne vrednosti označavaju da je proizvodnja električne energije veća od potrošnje.

Model/metrika	Udobnost (mean ± std)	Cena električne energije (mean ± std)	Ukupna nagrada (mean ± std)	Broj pokretanja uređaja (od 14)
DDQN	2,90 ± 3,10	-3,52 ± 0,34	-0,62 ± 2,92	12
PPO	5,22 ± 0,39	-3,59 ± 0,28	1,63 ± 0,10	14
KBS LLM	5,35 ± 0,55	-3,47 ± 0,40	1,87 ± 0,16	14
KBS definisan od strane eksperta	5,50 ± 0,44	-3,56 ± 0,32	1,92 ± 0,14	14

Rezultati modela su slični u dva različita simulaciona okruženja. DDQN model je imao najslabije rezultate, pri čemu nije uspešno aktivirao mašine zbog čega nisu upoređeni rezultati metrike cene potrošnje sa ostalim modelima.



Slika 72. Grafički prikaz performansi modela ostvarenih u *Load Flow* simulacionom okruženju u sedam uzastopnih testnih dana u januaru.



Slika 73. Grafički prikaz performansi modela ostvarenih u *Typhoon HIL* simulacionom okruženju u sedam uzastopnih testnih dana u januaru.

Tabela 24 prikazuje rezultate koje modeli postižu u *Load Flow* simulacionom okruženju u junu, dok slika 74 daje njihov grafički prikaz. KBS definisan od strane ljudskog eksperta ostvario je najstabilnije i najuspešnije rezultate samo u pogledu udobnosti. U pogledu ukupne nagrade, PPO model je ostvario najbolje rezultate, dok je KBS definisan od strane ljudskog eksperta bio sledeći najbliži sa 12% manjom ukupnom nagradom, dok su DDQN i KBS LLM imali slabije rezultate sa umanjnjem od 43% i 55%. U pogledu električne potrošnje, DDQN je ostvario najbolji rezultat, ali, pošto nije uspešno aktivirao jedan uređaj, možemo konstatovati da je PPO model u ovom pogledu postigao najbolji rezultat. Preostali modeli su ostvarili značajno gore rezultate u pogledu potrošnje električne energije. KBS LLM i KBS definisan od strane ljudskog eksperta su ostvarili 322% i 652% manju uštedu u odnosu na PPO model. U pogledu udobnosti, najbolje rezultate je ostvario KBS model definisan od strane ljudskog eksperta, gde su sledeći po redu PPO, KBS LLM i DDQN sa umanjnjem od 48%, 54% i 77%.

Tabela 24. *Load Flow* okruženje: srednje vrednosti (*mean*) i standardne devijacije (*std*) metrika udobnosti, cene električne energije i ukupne nagrade, koje modeli ostvaruju tokom sedam testnih dana u junu. Negativne vrednosti označavaju da je potrošnja energije bila veća od proizvodnje, dok pozitivne vrednosti označavaju da je proizvodnja električne energije veća od potrošnje.

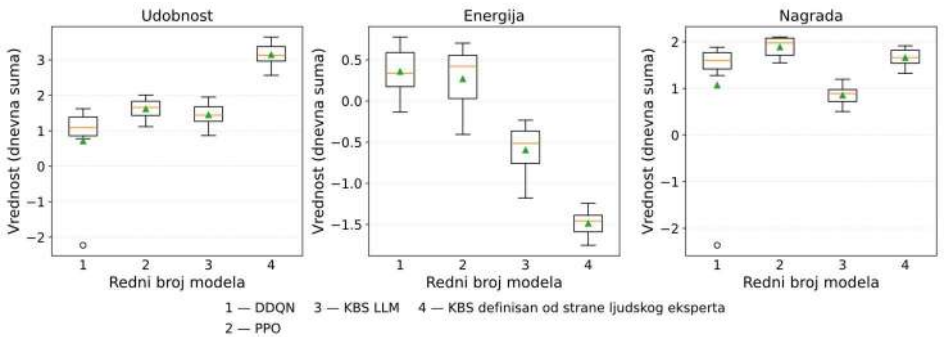
Model/ metrika	Udobnost (mean ± std)	Cena električne energije (mean ± std)	Ukupna nagrada (mean ± std)	Broj pokretanja uređaja (od 14)
DDQN	0,71 ± 1,23	0,36 ± 0,30	1,07 ± 1,41	13
PPO	1,62 ± 0,30	0,27 ± 0,37	1,89 ± 0,22	14
KBS LLM	1,45 ± 0,34	-0,60 ± 0,32	0,85 ± 0,22	14
KBS definisan od strane eksperta	3,15 ± 0,34	-1,49 ± 0,16	1,66 ± 0,20	14

Tabela 25 prikazuje rezultate koje modeli postižu u *Typhoon HIL* simulacionom okruženju u junu, dok slika 75 daje njihov grafički prikaz. KBS definisan od strane ljudskog eksperta ostvario je najstabilnije i najuspešnije rezultate samo u pogledu udobnosti, dok su ostali modeli PPO, DDQN i KBS LLM ostvarili umanjnje od 23%, 32% i 40%. U ostalim metrikama (ceni električne energije i ukupne nagrade) PPO model je ostvario najbolje rezultate. U pogledu ukupne nagrade DDQN, KBS LLM i KBS

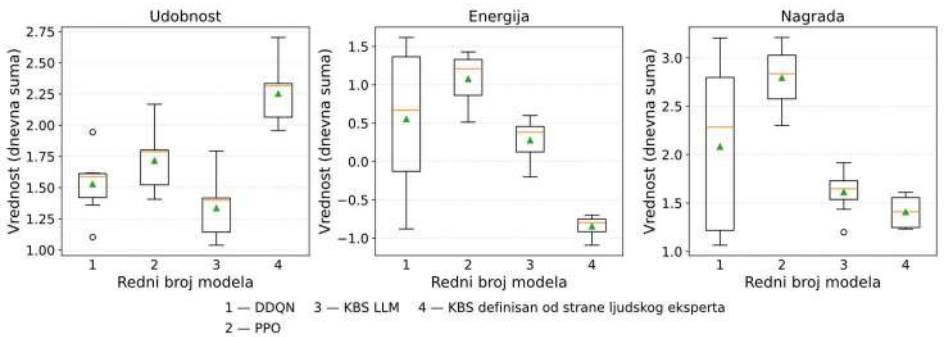
definisani od strane ljudskog eksperta su imali umanjene vrednosti za 25%, 43% i 50%. U pogledu cene električne energije su imali, u istom redosledu, umanjene vrednosti od 49%, 74%, i 178%. KBS definisan od strane ljudskog eksperta ima značajno umanjeno jer model veći značaj pridaje udobnosti.

Tabela 25. *Typhoon HIL* okruženje: srednje vrednosti i standardne devijacije metrika udobnosti, cene električne energije i ukupne nagrade, koje modeli ostvaruju tokom sedam testnih dana u junu. Negativne vrednosti označavaju da je potrošnja energije bila veća od proizvodnje, dok pozitivne vrednosti označavaju da je proizvodnja električne energije veća od potrošnje.

Model/metrika	Udobnost (mean ± std)	Cena električne energije (mean ± std)	Ukupna nagrada (mean ± std)	Broj pokretanja uređaja (od 14)
DDQN	1,53± 0,24	0,55 ± 0,91	2,10 ± 0,84	14
PPO	1,72 ± 0,24	1,08 ± 0,31	2,80 ± 0,30	14
KBS LLM	1,34 ± 0,24	0,28 ± 0,26	1,61 ± 0,22	14
KBS definisan od strane eksperta	2,25 ± 0,24	-0,85 ± 0,12	1,41 ± 0,16	14



Slika 74. Grafički prikaz performansi modela ostvarenih u *Load Flow* simulacionom okruženju u sedam uzastopnih testnih dana u junu.



Slika 75. Grafički prikaz performansi modela ostvarenih u *Typhoon HIL* simulacionom okruženju u sedam uzastopnih testnih dana u junu.

6.4.1 Zaključak ispitivanja robustnosti modela

U *Load Flow* simulacionom okruženju u januaru, najbolje rezultate u pogledu ukupne nagrade je ostvario KBS definisan od strane ljudskog eksperta. KBS LLM i PPO su postigli 15% i 23% nižu ukupnu nagradu. DDQN model nije uvek aktivirao uređaje u toku dana, zbog čega nije pouzdan za upotrebu u realnim uslovima.

U *Load Flow* simulacionom okruženju u junu, najbolje rezultate u pogledu ukupne nagrade je ostvario PPO model, dok je KBS definisan od strane ljudskog eksperta ostvario 12% manju ukupnu nagradu. DDQN i KBS LLM su imali znatno slabije performanse u odnosu na PPO (43% i 55% manju ukupnu nagradu). Ponovo, DDQN model nije uvek uspešno aktivirao sve uređaje.

U *Typhoon HIL* simulacionom okruženju su rezultati bili slični. U januaru je najbolje rezultate ostvario KBS definisan od strane ljudskog eksperta, a KBS LLM i PPO su imali ukupnu nagradu nižu za 3% i 15%. I u ovom slučaju, DDQN model nije uvek uspešno aktivirao uređaje. U junu, PPO je postigao najbolje rezultate u pogledu ukupne nagrade. KBS definisan od strane ljudskog stručnjaka je ostvario 50% nižu ukupnu nagradu.

Rezultati su se značajno razlikovali između *Load Flow* i *Typhoon HIL* simulacionih okruženja zbog načina simulacije potrošnje uređaja. Kod *Load Flow* simulacionog okruženja potrošnja električne energije pojedinačnog uređaja varira u toku njegovog rada. Ova dinamičnost povećava kompleksnost optimizacije modela ali i omogućava finije upravljanje uređajima. PPO model je u ovom slučaju pokazao veću fleksibilnost i sposobnost pronalaženja optimalnog kompromisa između cene električne energije i udobnosti stanara.

Sa druge strane, u *Typhoon HIL* simulacionom okruženju, uređaji su simulirani sa konstantnom potrošnjom tokom rada, što pojednostavljuje proces učenja jer modeli imaju stabilnije ulazne obrasce, ali je takva simulacija manje realistična.

Posmatrajući rezultate u pogledu oba simulaciona okruženja i oba godišnja doba, KBS definisan od strane ljudskog eksperta se pokazao najbolji u zimskom periodu, dok se PPO model pokazao najbolji u letnjem periodu. PPO model je u više scenarija, posebno u junu, nadmašio ostale modele u pogledu ukupne nagrade i cene električne energije. Ovim je postignuta automatska prilagodljivost što predstavlja ključnu prednost RL modela u odnosu na statički definisane ekspertske sisteme, jer ovaj pristup omogućava efikasno reagovanje na različite sezonske i dinamične promene u potrošnji. Prednost RL pristupa je njihova potpuna automatizacija, zahvaljujući kojoj stanar nije opterećen ručnim podešavanjem rasporeda rada uređaja, već sistem sam uči optimalne obrasce ponašanja.

U januaru KBS LLM je ostvario veće vrednosti ukupne nagrade u odnosu na PPO u oba simulaciona okruženja, sa poboljšanjem od 10% u *Load Flow* simulacionom okruženju i 14% u *Typhoon HIL* simulacionom okruženju. U junu je PPO ostvario 122% veću ukupnu nagradu u *Load Flow* simulacionom okruženju i 74% u *Typhoon HIL* simulacionom okruženju. Iako je KBS LLM ostvario gore performanse u odnosu na PPO model, prednost LLM modela je sposobnost automatskog prilagođavanja promenama u okruženju bez potrebe za ponovnim treniranjem ili dodatnim finim podešavanjem.

Shuvo i Yilmaz (2022) su predložili model zasnovan na akter-kritičar algoritmu (eng. *Advantage Actor-Critic*), koji upravlja radom kućnih uređajima sa ciljem optimizacije potrošnje električne energije i očuvanja udobnosti stanara u pametnoj kući. Kao

referentni model autori koriste pristup zasnovan na ručnom upravljanju kućnim uređajima, u kojem je ponašanje modela direktno kontrolisano od strane stanara, bez primene automatizovanog planiranja ili optimizacije. Okruženje za simulaciju je modelovalo uređaje sa varijabilnom potrošnjom (klima uređaj), konstantnom potrošnjom (mašina za veš i sudove) i kombinovanom potrošnjom (električno vozilo). Akter-kritičar algoritam je ostvario 23% bolju funkciju nagrade u odnosu na referentni model. Međutim, treba istaći da je evaluacija sprovedena na istom periodu koji je korišćen za treniranje modela, odnosno, nije izdvojen poseban test skup kao u ovoj disertaciji. Takođe, razmatrani period je obuhvatao 30 uzastopnih dana u prolećnom periodu, što znači da nije ispitivan uticaj sezonskih promena. Ova teza unapređuje pomenutu studiju proširivanjem eksperimentalne postavke na različita simulaciona okruženja, različite tipove potrošnje (konstantnu i varijabilnu) i analizu uticaja sezonalnosti. PPO model je u ovoj disertaciji ostvario značajno bolje rezultate u letnjem periodu u poređenju sa referentnim KBS definisanim od strane ljudskog eksperta (za 98% u *Typhoon HIL* simulacionom okruženju i za 14% u *Load Flow* simulacionom okruženju), gde je evaluacija odrađena na sedam testnih dana.

Li i saradnici (Li et al. 2020) predložili su PPO zasnovani pristup za upravljanje kućnim energetskim sistemom sa ciljem minimizacije troškova električne energije u uslovima neizvesnih cena i ponašanja stanara. Sistem obuhvata kritične, prilagodljive i kontrolisane uređaje, pri čemu se funkcija nagrade zasniva na trošku električne energije i penalizaciji nedovoljne napunjenosti baterije električnog vozila. Rezultati pokazuju da PPO model ostvaruje smanjenje troškova od oko 40% u odnosu na referentni model bez automatizovanog upravljanja, dok teoretski optimalni model postiže približno 3% niže troškove u poređenju sa PPO modelom. Evaluacija je sprovedena u stohastički modelovanom simulacionom okruženju, bez analize komfora stanara i sezonskih uticaja. Nasuprot tome, u okviru ove teze razmatran je referentni KBS dizajniran od strane ljudskog eksperta sa ciljem optimizacije koji kombinuje potrošnju energije sa udobnošću stanara.

Rad (Lissa et al. 2021) je predložio strategiju učenja potkrepljivanjem sa ciljem rešavanja problema upravljanja unutrašnjom temperaturom i radom toplotne pumpe. Treniranje i evaluacija sprovedeni su u periodu od maja do decembra, pri čemu je simulirana varijabilna potrošnja energije od strane uređaja. Autori su definisali dva referentna modela. Prvi referentni model je bio jednostavan, zasnovan na fiksnim pravilima upravljanja, sa ciljem održavanja unutrašnje temperature unutar definisanog opsega temperaturnih granica, bez korišćenja informacija o PV proizvodnji električne energije. Drugi referentni model je takođe bio zasnovan na pravilima, ali je umesto fiksnih koristio dinamički definisane temperaturne granice, u zavisnosti od doba dana, čime je omogućeno fleksibilnije upravljanje komforom kroz dozvoljena odstupanja od nominalne temperature. U poređenju sa referentnim modelima, predloženi DQN model je smanjio potrošnju električne energije za 2,66% u junu, ali je u decembru rezultovao 0.26% povećanom potrošnjom električne energije. Autori su definisali udobnost kao procenat preklapanja unutrašnje temperature sa željenim temperaturnim opsegom, koji je za sve modele iznosio oko 99%.

U ovoj tezi, u pogledu cene električne energije, PPO model je u oba simulaciona okruženja u januaru i junu postigao bolje rezultate cene električne energije u odnosu na referentni model, osim u jednoj postavci. U *Typhoon HIL* simulacionom okruženju u junu rezultati su bili bolji za 227%, dok su u *Load Flow* simulacionom okruženju bili bolji za

118%. U januaru, u *Typhoon HIL* simulacionom okruženju rezultati cene električne energije PPO modela bili su slabiji za 0,84%, dok su u *Load Flow* simulacionom okruženju bili bolji za 3,5%. Značajno veća ušteda električne energije ukazuje na to da je predloženi PPO model efikasnije rešenje za složenija simulaciona okruženja sa većim brojem uređaja i promenljivom PV proizvodnjom.

6.5 Zaključci i potencijalna poboljšanja

U ovoj disertaciji evaluirana je hipoteza da su metode zasnovane na učenju potkrepljivanjem i veliki jezički modeli efikasni pristupi za rešavanje postavljenog optimizacionog problema balansiranja potrošnje električne energije i udobnosti stanara. Efikasnost predloženih pristupa je definisana kroz sledeće kriterijume:

- (1) Predloženo rešenje postiže iste ili bolje vrednosti funkcije nagrade u odnosu na KBS definisan od strane ljudskog eksperta i
- (2) Adekvatnost odluka iz perspektive ljudskog eksperta.

Posmatrajući vrednost ukupne nagrade, PPO model ne ostvaruje bolje rezultate u svim razmatranim slučajevima u poređenju sa KBS definisanim modelom od strane ljudskog eksperta. U zimskom periodu PPO model ostvario je 23% i 15% manju ukupnu nagradu u *Load Flow* i *Typhoon HIL* simulacionom okruženju, tim redosledom. Nasuprot tome, u letnjem periodu, PPO model je ostvario značajno bolje rezultate, sa ukupnom većom za 14% i 98%, u u *Load Flow* i *Typhoon HIL* simulacionom okruženju, tim redosledom. Za razliku od ručno dizajniranih sistema zasnovanih na znanju, koji zahtevaju eksplicitno redefinisane pravila pri svakoj promeni uslova rada, PPO model omogućava adaptivno učenje politike direktno iz interakcije sa okruženjem, što ga čini pogodnijim za rad u okruženjima sa nepoznatom ili promenljivom dinamikom. Posmatrano na osnovu ukupnih rezultata, može se zaključiti da PPO model ispunjava prvi kriterijum. Rezultati takođe ukazuju na potrebu prilagođavanja modela sezonskim uslovima i karakteristikama simulacionog okruženja, što je PPO model ostvario kroz proces ponovnog treniranja. U literaturi je zapaženo da je modele sa eksplicitno definisanim ciljem optimizacije, uključujući ljudski definisane KBS, teško nadmašiti u svim uslovima rada (Lissa et al. 2021), (Li et al. 2020), (Mbuwir et al. 2021) i (Aldahmashi and Ma 2024). Kada referentni model ne uključuje automatizovano planiranje i raspoređivanje rada kućnih uređaja, niti eksplicitno definisan cilj optimizacije potrošnje električne energije, razlika u performansama između predloženog algoritma i referentnog pristupa postaje izrazito velika (Li et al. 2020), (Mbuwir et al. 2021) i (Aldahmashi and Ma 2024). Budući da je KBS dizajniran od strane ljudskog eksperta u ovoj disertaciji sadržao jasno definisan cilj optimizacije, manja razlika u ukupnoj nagradi u odnosu na predloženo rešenje može se smatrati očekivanom.

Za razliku od PPO modela, DDQN model je u svim razmatranim postavkama ostvario lošije performanse u poređenju sa KBS modelom definisanim od strane ljudskog eksperta. Pored znatno nižih vrednosti ukupne nagrade, uočeno je i nestabilno ponašanje modela, gde je DDQN model u pojedinim scenarijama propustio da pokrene sve uređaje. Na osnovu ovih rezultata, možemo zaključiti da DDQN model ne zadovoljava prvi kriterijum uspeha.

U pogledu ukupne nagrade, KBS LLM model u gotovo svim razmatranim slučajevima ne ostvaruje bolje rezultate u poređenju sa KBS modelom definisanim od strane ljudskog

eksperta. U januaru je KBS LLM ostvario 15% i 3% nižu ukupnu nagradu u *Load Flow* i *Typhoon HIL* simulacionom okruženju, tim redosledom. U letnjem periodu je razlika u ukupnoj nagradi bila znatno izraženija. KBS LLM je ostvario 49% manju ukupnu nagradu u *Load Flow* simulacionom okruženju, dok je u *Typhoon HIL* postigao 14% veću ukupnu nagradu. Ipak, ključna prednost KBS LLM nad ljudski definisanim KBS je što ne zahteva ručni dizajn pravila, koji bi morao da se ponavlja pri svakoj promeni okruženja. Dodatno, u poređenju sa RL modelima, KBS LLM eliminiše potrebu za dugotrajnim procesom treniranja. Na osnovu ukupnih rezultata može se zaključiti da KBS LLM model ne ispunjava prvi postavljeni kriterijum, ali ipak predstavlja obećavajući pravac daljih istraživanja.

Usled ograničenja resursa, RL LLM je mogao biti pokrenut samo na jednom testnom danu u januaru i u junu, u *Load Flow* simulacionom okruženju. U datom slučaju, RL LLM model nije ostvario bolje rezultate u poređenju sa KBS modelom definisanim od strane ljudskog eksperta. U zimskom danu, RL LLM je ostvario 11% nižu ukupnu nagradu, dok je u letnjem ostvario 44% nižu ukupnu nagradu. Slično KBS LLM pristupu, RL LLM ne zahteva eksplicitno definisanje pravila od strane ljudskog eksperta, niti dugotrajan proces obučavanja kao RL modeli. Međutim, RL LLM model zahteva kontinuiranu komunikaciju sa velikim jezičkim modelom tokom izvršavanja, što ga čini skupljim i infrastrukturno zahtevnijim za primenu u odnosu na KBS LLM. Zbog ovih ograničenja, rezultati su evaluirani isključivo na nivou jednog dana i u *Load Flow* simulacionom okruženju. Na osnovu rezultata može se zaključiti da RL LLM model ne nadmašuje ljudski definisan KBS po kriterijumu ukupne nagrade.

Posmatrajući iznesene rezultate, PPO model se pokazao kao najefikasniji pristup, posebno u uslovima koji zahtevaju prilagođavanje sezonskim promenama i karakteristikama simulacionog okruženja. Iako u pojedinim scenarijima ne nadmašuje KBS model definisan od strane ljudskog eksperta, PPO model ostvaruje konkurentne ili bolje rezultate bez značajnog pogoršanja performansi, čime ispunjava prvi kriterijum efikasnosti. Sa druge strane, LLM-generisana rešenja pokazuju značajan potencijal. Posebno se KBS LLM ističe kao perspektivna praktična alternativa KBS modelu definisanom od strane ljudskog eksperta. Iako RL LLM ne nadmašuje ljudski definisan model po kriterijumu ukupne nagrade, postignuti rezultati ukazuju na izvodljivost integracije velikih jezičkih modela u procesu donošenja odluka.

U pogledu drugog kriterijuma, analiza sprovedena u potpoglavlju 6.1.3 je pokazala je da su odluke PPO modela razumne iz perspektive ljudskog eksperta. SHAP analiza (potpoglavlje 6.1.4) je pokazala da se PPO model pri donošenju odluka oslanja na relevantne karakteristike stanja, te da su odluke PPO modela povezane sa stvarnim sezonskim i tarifnim uslovima. Dakle, možemo zaključiti da PPO model ispunjava i drugi kriterijum uspeha.

Sa druge strane, analiza odluka DDQN modela je pokazala da odluke modela većinom zavise od tarife, dok se druge relevantne karakteristike zanemaruju. Analiza je pokazala i da DDQN model u određenim situacijama ne upravlja uređajima na optimalan način, kao i da povremeno propušta da aktivira uređaj tokom dana. Stoga možemo zaključiti da DDQN model ne zadovoljava drugi kriterijum uspeha.

Uzimajući u obzir sve analizirane metrike, postavljena hipoteza je potvrđena za PPO model. Iako RL LLM pristup primenjen u ovom radu ne ispunjava postavljene

kriterijume uspeha, ostvareni rezultati pokazuju da je ovaj pravac istraživanja obećavajući. Bolji rezultati bi se mogli postići unapređenjem dizajna prompt-ova, pažljivim izborom i strukturiranjem primera za *few-shot* pristup, kao i prilagođavanje velikog jezičkog modela specifičnostima razmatranog zadatka kroz postupke finog podešavanja (eng. *fine-tuning*).

7 Zaključak

U ovoj disertaciji razmatra se problem automatskog upravljanja potrošnjom električne energije u okviru jednog domaćinstva, sa ciljem balansiranja troškova električne energije i komfora stanara, uz integraciju obnovljivih izvora energije. Motivacija za rešavanje ovog problema proizlazi iz rastuće globalne potražnje za energijom i potrebe za smanjenjem emisija gasova sa efektom staklene bašte, pri čemu pametne kuće predstavljaju značajnu priliku za unapređenje energetske efikasnosti na lokalnom nivou. Primena metoda dubokog učenja, a posebno dubokog učenja potkrepljivanjem, omogućava adaptivnu automatizaciju rada kućnih uređaja u okruženjima sa promenljivom dinamikom, čime se postiže efikasnije korišćenje električne energije uz očuvanje kvaliteta stanovanja.

Cilj ove teze je ispitivanje efikasnosti savremenih metoda veštačke inteligencije u rešavanju optimizacionog problema upravljanja potrošnjom električne energije i komforom stanara u okviru jednog domaćinstva. U tom kontekstu razmatrana su dva pristupa: metode učenja potkrepljivanjem i pristupi zasnovani na velikim jezičkim modelima. Istraživačka hipoteza teze glasi da oba pristupa mogu ostvariti iste ili bolje vrednosti funkcije nagrade u poređenju sa manuelno definisanim sistemom zasnovanim na znanju, pri čemu su odluke koje donose modeli adekvatne iz perspektive ljudskog eksperta, što se potvrđuje kroz analizu objašnjivosti modela.

Radi evaluacije i treniranja modela razvijeno je simulaciono okruženje koje verno oponaša potrošnju električne energije u pametnoj kući. Okruženje je dizajnirano da bude adaptivno, odnosno na taj način da je lako dodavati različite tipove uređaja i definisati raznolike scenarije u pogledu spoljnih vremenskih uslova i tarifnih režima. U pogledu simulacije rada uređaja, korišćena su dva pristupa. U *Typhoon HIL* okruženju, uređaji su modelovani tako da imaju konstantnu potrošnju tokom rada, dok su u *Load Flow* okruženju modelovani tako da imaju varijabilnu potrošnju u toku rada, što više odgovara realnoj situaciji.

Radi evaluacije predloženog pristupa definisana je višeciljna funkcija nagrade koja objedinjuje cenu električne energije i komfor stanara. Metrika komfora stanara izražena je u monetarnim jedinicama, čime je omogućeno njeno direktno sabiranje sa troškovima električne energije unutar funkcije nagrade. Ova metrika kvantifikuje spremnost stanara da plate uslugu rada uređaja u određenom periodu dana. Konkretno vrednosti određene su uvođenjem tri nivoa prioriteta uređaja (visok, srednji i nizak), kao i njihovim mapiranjem na odgovarajuće tarifne režime električne energije (visoka, srednja i niska tarifa).

Dizajnirana su sledeća rešenja za automatizaciju planiranja vremena rada uređaja u pametnoj kući, s ciljem postizanja maksimalne definisane nagrade:

- Metode zasnovane na učenju potkrepljivanjem: DDQN i PPO modeli.
- Pristupi zasnovani na primeni velikih jezičkih modela:
 - KBS LLM: LLM-u je prosleđena instrukcija da, na osnovu opisa okruženja i zahteva stanara, generiše sistem zasnovan na pravilima.

- RL LLM: LLM je tretiran kao model učenja potkrepljivanjem, pri čemu je u svakom promptu prosleđivan opis trenutnog stanja na osnovu kog je birao optimalnu akciju. Dodatno, na raspolaganju su mu bili primeri prethodnih interakcija, sa dostupnim vrednostima neposredne nagrade za ranija stanja i sprovedene akcije.

Radi ispitivanja hipoteze disertacije, dizajnirana rešenja upoređena su sa referentnim modelom, koji predstavlja KBS sistem dizajniran od strane ljudskog eksperta. Ponašanje predloženih rešenja analizirano je u odnosu na referentni model u različitim uslovima, u pogledu simulacionog okruženja i godišnjeg doba. Pored poređenja ukupne vrednosti funkcije nagrade, sprovedena je i detaljna analiza odluka modela, koja je obuhvatila ručnu analizu ponašanja, kao i SHAP analizu doprinosa pojedinačnih karakteristika stanja u formiranju konačnih predikcija obučениh modela.

Rezultati su pokazali da se PPO model obučava stabilnije i brže u odnosu na DDQN model u svim razmatranim scenarijima. Takođe je utvrđeno da PPO model ispunjava postavljene kriterijume: posmatrano globalno, kroz sve testirane scenarije, ostvaruje veću vrednost ukupne funkcije nagrade u poređenju sa KBS sistemom definisanim od strane ljudskog eksperta, dok su odluke koje donosi razumljive iz perspektive ljudskog eksperta i zasnovane na relevantnim karakteristikama stanja. Nasuprot tome, DDQN model nije ispunio nijedan od definisanih kriterijuma.

Pristupi zasnovani na velikim jezičkim modelima nisu ispunili kriterijum prema kojem je bilo očekivano da ostvare bolje rezultate u odnosu na referentni model. Ipak, rezultati su pokazali da performanse sistema u velikoj meri zavise od godišnjeg doba i korišćenog simulacionog okruženja. Prednost svih predloženih modela u odnosu na referentni model ogleda se u njihovoj sposobnosti automatske adaptacije na ove uslove, dok bi KBS sistem definisan od strane ljudskog eksperta zahtevao redizajn pri svakoj promeni okruženja.

Imajući u vidu da pristupi zasnovani na velikim jezičkim modelima nisu ostvarili značajno lošije rezultate u poređenju sa referentnim modelom, ovaj pravac istraživanja može se smatrati obećavajućim. Takođe, važno je naglasiti da su u okviru ove disertacije razmatrani osnovni oblici ovih pristupa, te da bi se njihove performanse potencijalno mogle unaprediti unapređenjem dizajna promptova, pažljivijim izborom i strukturiranjem primera za RL LLM pristup, kao i prilagođavanjem velikog jezičkog modela specifičnostima razmatranog zadatka kroz postupke finog podešavanja.

Na osnovu dobijenih rezultata, mogu se izdvojiti dodatni pravci daljeg unapređenja predloženih rešenja. S obzirom na to da PPO model nije nadmašio referentni model u svim razmatranim scenarijima, jedan od potencijalnih pravaca predstavlja hibridni pristup koji kombinuje PPO model i KBS sistem definisan od strane ljudskog eksperta, čime bi se objedinile prednosti učenja iz podataka i ekspertskog znanja.

Dodatni pravac unapređenja odnosi se na unapređenje reprezentacije stanja sistema. U okviru ovog istraživanja stanje je definisano na osnovu trenutnih merenja i poznatih parametara okruženja, što predstavlja uobičajen i praktičan pristup, ali ne uključuje eksplicitne informacije o budućim promenama u potrošnji i proizvodnji električne energije. U tom smislu, potencijalno unapređenje podrazumevalo bi proširenje prostora stanja uključivanjem predikcija eksternih modela, kao što su kratkoročne prognoze PV proizvodnje i potrošnje pojedinačnih uređaja.

Evaluaciona postavka mogla bi se dodatno unaprediti primenom modela na podatke koji obuhvataju celu godinu, uz korišćenje pristupa zasnovanog na pokretnom vremenskom prozoru. U ovom pristupu modeli bi se periodično ponovo obučavali na proširenom skupu podataka, koji uključuje najnovije dostupne informacije, čime bi se omogućila kontinuirana adaptacija na sezonske promene i dugoročne obrasce u potrošnji i proizvodnji električne energije.

Testirani scenariji mogli bi se dodatno obogatiti omogućavanjem prodaje viška proizvedene električne energije električnoj mreži, kao i uvođenjem baterijskog sistema za skladištenje energije u pametnoj kući. Konačno, nivo optimizacije mogao bi se proširiti sa nivoa pojedinačne kuće na nivo zgrade ili čak celog naselja. Ovakav pristup omogućio bi efikasniju upotrebu električne energije, pri čemu bi se višeciljna optimizacija primenjivala na širem planu, čime bi se automatizovalo upravljanje energetskim resursima u većim razmerama.

U okviru ove disertacije dizajnirana je ciljna funkcija nagrade koja je formalno usklađena sa željama korisnika i istovremeno pogodna za treniranje modela učenja potkrepljivanjem. Međutim, ostaje otvoreno pitanje u kojoj meri ovako definisana funkcija nagrade verno odražava stvarne preferencije korisnika, naročito u pogledu kompromisa između troškova električne energije i subjektivnog komfora. Iako je komfor modelovan kroz unapred definisane prioritete i monetarne ekvivalente, ovakav pristup neminovno predstavlja aproksimaciju složenih i često kontekstualno zavisnih korisničkih preferencija. Zbog toga bi adekvatniji postupak evaluacije podrazumevao uključivanje korisnika u proces testiranja, kroz prikupljanje eksplicitnih povratnih informacija o donetim odlukama sistema, koje bi se mogle koristiti za dodatno prilagođavanje ciljne funkcije ili politike upravljanja.

Važno je takođe istaći da su sva razmatrana rešenja evaluirana u simulacionom okruženju, te je neophodno ispitati u kojoj meri bi se njihovo ponašanje prenelo na realne uslove rada pametne kuće. Jedan od praktičnih pristupa podrazumevao bi treniranje modela u simulaciji, a zatim primenu već obučenog modela u stvarnom okruženju, uz mogućnost njegovog dodatnog prilagođavanja na osnovu realnih podataka.

Upravo u tom smislu, očekivani rezultati ovog istraživanja imaju značajnu praktičnu i naučnu vrednost. Definisani simulator pametne kuće, zajedno sa komunikacionim slojem ka modelu upravljanja, predstavlja osnovu adaptivnog okruženja sposobnog da simulira proizvoljne konfiguracije pametnih kuća. Takav simulator omogućava realistične scenarije za obučavanje i evaluaciju modela pre njihove primene u stvarnim uslovima, čime se dodatno smanjuje rizik od neželjenog ponašanja sistema, koje bi moglo dovesti do nepotrebnih troškova električne energije ili narušavanja navika i komfora stanara. Dodatnu praktičnu vrednost predstavlja definisana funkcija nagrade, koja omogućava stanarima da na intuitivan način iskažu svoje potrebe i prioritete. Modeli obučeni u simulaciji, koji su pokazali sposobnost donošenja optimalnih odluka, mogu stoga poslužiti kao polazna osnova za implementaciju sistema upravljanja energijom (HEMS) u stvarnoj pametnoj kući.

Naučni doprinos ovog istraživanja ogleda se u tome što simulator i definisana funkcija nagrade predstavljaju reproduktivnu i proširivu istraživačku platformu, koju drugi istraživači mogu koristiti za razvoj i evaluaciju sopstvenih HEMS rešenja. Pored toga, ovo istraživanje spada među prva koja sistematski ispituju primenu pretreniranih velikih

jezičkih modela u kontekstu automatizovanog upravljanja energijom u pametnim kućama.

Literatura

- Shapley, Lloyd S. "A value for n-person games." (1953): 307-317.
- Agnetis, Alessandro, Gianluca De Pascale, Paolo Detti, and Antonio Vicino. 2013. "Load Scheduling for Household Energy Consumption Optimization." *IEEE Transactions on Smart Grid* 4 (4): 2364–73. <https://doi.org/10.1109/TSG.2013.2254506>.
- Ahmad, Adnan, Asif Khan, Nadeem Javaid, et al. 2017. "An Optimized Home Energy Management System with Integrated Renewable Energy and Storage Resources." *Energies* 10 (4): 549. <https://doi.org/10.3390/en10040549>.
- Al-Ani, Omar, and Sanjoy Das. 2022. "Reinforcement Learning: Theory and Applications in HEMS." *Energies* 15 (17): 6392. <https://doi.org/10.3390/en15176392>.
- Aldahmashi, Jamal, and Xiandong Ma. 2024. "Real-Time Energy Management in Smart Homes Through Deep Reinforcement Learning." *IEEE Access* 12: 43155–72. <https://doi.org/10.1109/ACCESS.2024.3375771>.
- Allegretti, G., M.A. Montoya, L.A.S. Bertussi, and E. Talamini. 2022. "When Being Renewable May Not Be Enough: Typologies of Trends in Energy and Carbon Footprint towards Sustainable Development." *Renewable and Sustainable Energy Reviews* 168 (October): 112860. <https://doi.org/10.1016/j.rser.2022.112860>.
- M. Almashor and Y. Miyashita, "Can Private LLM Agents Synthesize Household Energy Consumption Data?," in Proc. 15th ACM Int. Conf. Future and Sustainable Energy Systems, pp. 664–668, Jun. 2024, doi: 10.1145/3632775.3661993.
- Arulkumaran, Kai, Marc Peter Deisenroth, Miles Brundage, and Anil Anthony Bharath. 2017. "A Brief Survey of Deep Reinforcement Learning." *IEEE Signal Processing Magazine* 34 (6): 26–38. <https://doi.org/10.1109/MSP.2017.2743240>.
- Azuatalam, Donald, Wee-Lih Lee, Frits de Nijs, and Ariel Liebman. 2020. "Reinforcement Learning for Whole-Building HVAC Control and Demand Response." *Energy and AI* 2 (November): 100020. <https://doi.org/10.1016/j.egyai.2020.100020>.
- Badini, Silvia, Stefano Regondi, Emanuele Frontoni, and Raffaele Pugliese. 2023. "Assessing the Capabilities of ChatGPT to Improve Additive Manufacturing Troubleshooting." *Advanced Industrial and Engineering Polymer Research* 6 (3): 278–87. <https://doi.org/10.1016/j.aiepr.2023.03.003>.
- Barbato, Antimo, Luca Borsani, Antonio Capone, and Stefano Melzi. 2009. "Home Energy Saving through a User Profiling System Based on Wireless Sensors." *Proceedings of the First ACM Workshop on Embedded Sensing Systems for Energy-Efficiency in Buildings*, November 3, 49–54. <https://doi.org/10.1145/1810279.1810291>.
- Barbier, Edward B., and Joanne C. Burgess. 2017. "The Sustainable Development Goals and the Systems Approach to Sustainability." *Economics* 11 (1): 20170028. <https://doi.org/10.5018/economics-ejournal.ja.2017-28>.
- Beechey, Daniel, Thomas M. S. Smith, and Özgür Şimşek. 2023. "Explaining Reinforcement Learning with Shapley Values." *Proceedings of the 40th International Conference on Machine Learning*, July 3, 2003–14. <https://proceedings.mlr.press/v202/beeche23a.html>.

- Bengio, Yoshua, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. "A neural probabilistic language model." *Journal of machine learning research* 3, no. Feb (2003): 1137-1155.
- Daniel, Bick. "Towards Delivering a Coherent Self-Contained Explanation of Proximal Policy Optimization." PhD diss., M. Sc. Thesis, University of Groningen, The Netherlands, 2021.
- Boynuegri, A.R., Bunyamin Yagcitekin, Mustafa Baysal, Arif Karakas, and Mehmet Uzunoglu. 2013. "Energy Management Algorithm for Smart Home with Renewable Energy Sources." May 1, 1753–58. <https://doi.org/10.1109/PowerEng.2013.6635883>.
- Brown, Tom, Benjamin Mann, Nick Ryder, et al. 2020. "Language Models Are Few-Shot Learners." *Advances in Neural Information Processing Systems* 33: 1877–901. https://proceedings.neurips.cc/paper_files/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.
- Buratti, C., M. Vergoni, and D. Palladino. 2015. "Thermal Comfort Evaluation Within Non-Residential Environments: Development of Artificial Neural Network by Using the Adaptive Approach Data." *Energy Procedia* 78 (November): 2875–80. <https://doi.org/10.1016/j.egypro.2015.11.656>.
- Buren, David Van. 2023. "Guided Scenarios with Simulated Expert Personae: A Remarkable Strategy to Perform Cognitive Work." arXiv:2306.03104. Preprint, arXiv, June 3. <https://doi.org/10.48550/arXiv.2306.03104>.
- Cao, Yuji, Huan Zhao, Yuheng Cheng, et al. 2024. "Survey on Large Language Model-Enhanced Reinforcement Learning: Concept, Taxonomy, and Methods." arXiv:2404.00282. Preprint, arXiv, March 30. <http://arxiv.org/abs/2404.00282>.
- Chehri, Abdellah, and Hussein T. Mouftah. 2013. "FEMAN: Fuzzy-Based Energy Management System for Green Houses Using Hybrid Grid Solar Power." *Journal of Renewable Energy* 2013 (September): e785636. <https://doi.org/10.1155/2013/785636>.
- Chellaswamy, C., R. Ganesh Babu, and A. Vanathi. 2021. "A Framework for Building Energy Management System with Residence Mounted Photovoltaic." *Building Simulation* 14 (4): 1031–46. <https://doi.org/10.1007/s12273-020-0735-x>.
- Cobbe, Karl, Oleg Klimov, Chris Hesse, Taehoon Kim, and John Schulman. 2019. "Quantifying Generalization in Reinforcement Learning." *Proceedings of the 36th International Conference on Machine Learning*, May 24, 1282–89. <https://proceedings.mlr.press/v97/cobbe19a.html>.
- Collins, Michael. 1996. "A New Statistical Parser Based on Bigram Lexical Dependencies." arXiv:cmp-Lg/9605012. Preprint, arXiv, May 6. <https://doi.org/10.48550/arXiv.cmp-lg/9605012>.
- Cook, D.J., M. Youngblood, E.O. Heierman, et al. 2003. "MavHome: An Agent-Based Smart Home." *Proceedings of the First IEEE International Conference on Pervasive Computing and Communications, 2003. (PerCom 2003).*, 521–24. <https://doi.org/10.1109/PERCOM.2003.1192783>.
- Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding." In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, edited by Jill Burstein, Christy Doran, and Tamar Solorio. Association for Computational Linguistics. <https://doi.org/10.18653/v1/N19-1423>.

- Efrat, Avia, and Omer Levy. 2020. "The Turking Test: Can Language Models Understand Instructions?" arXiv:2010.11982. Preprint, arXiv, October 22. <https://doi.org/10.48550/arXiv.2010.11982>.
- Engelhardt, Raphael C., Moritz Lange, Laurenz Wiskott, and Wolfgang Konen. 2024. "Exploring the Reliability of SHAP Values in Reinforcement Learning." In *Explainable Artificial Intelligence*, edited by Luca Longo, Sebastian Lapuschkin, and Christin Seifert, vol. 2155. Communications in Computer and Information Science. Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-63800-8_9.
- Esmacel Nezhad, Ali, Abolfazl Rahimnejad, and S. Andrew Gadsden. 2021. "Home Energy Management System for Smart Buildings with Inverter-Based Air Conditioning System." *International Journal of Electrical Power & Energy Systems* 133 (December): 107230. <https://doi.org/10.1016/j.ijepes.2021.107230>.
- Fenner, Andriel Evandro, Charles Joseph Kibert, Junghoon Woo, et al. 2018. "The Carbon Footprint of Buildings: A Review of Methodologies and Applications." *Renewable and Sustainable Energy Reviews* 94 (October): 1142–52. <https://doi.org/10.1016/j.rser.2018.07.012>.
- Forootani, Ali, Mohammad Rastegar, and Mohammad Jooshaki. 2022. "An Advanced Satisfaction-Based Home Energy Management System Using Deep Reinforcement Learning." *IEEE Access* 10: 47896–905. <https://doi.org/10.1109/ACCESS.2022.3172327>.
- Fu, Qiming, Zhicong Han, Jianping Chen, You Lu, Hongjie Wu, and Yunzhe Wang. 2022. "Applications of Reinforcement Learning for Building Energy Efficiency Control: A Review." *Journal of Building Engineering* 50 (June): 104165. <https://doi.org/10.1016/j.jobbe.2022.104165>.
- Fuente, Neil De La, and Daniel A. Vidal Guerra. 2024. "A Comparative Study of Deep Reinforcement Learning Models: DQN vs PPO vs A2C." arXiv:2407.14151. Version 1. Preprint, arXiv, July 19. <https://doi.org/10.48550/arXiv.2407.14151>.
- Fui-Hoon Nah, Fiona, Ruilin Zheng, Jingyuan Cai, Keng Siau, and Langtao Chen. 2023. "Generative AI and ChatGPT: Applications, Challenges, and AI-Human Collaboration." *Journal of Information Technology Case and Application Research* 25 (3): 277–304. <https://doi.org/10.1080/15228053.2023.2233814>.
- Gao, Chen, Xiaochong Lan, Nian Li, et al. 2023. "Large Language Models Empowered Agent-Based Modeling and Simulation: A Survey and Perspectives." arXiv:2312.11970. Preprint, arXiv, December 19. <http://arxiv.org/abs/2312.11970>.
- Gao, Jianfeng, and Chin-Yew Lin. 2004. "Introduction to the Special Issue on Statistical Language Modeling." *ACM Transactions on Asian Language Information Processing* 3 (2): 87–93. <https://doi.org/10.1145/1034780.1034781>.
- Gehring, Jonas, Michael Auli, David Grangier, Denis Yarats, and Yann N. Dauphin. 2017. "Convolutional Sequence to Sequence Learning." *Proceedings of the 34th International Conference on Machine Learning*, July 17, 1243–52. <https://proceedings.mlr.press/v70/gehring17a.html>.
- Gholami, M., and M.J. Sanjari. 2021. "Multiobjective Energy Management in Battery-Integrated Home Energy Systems." *Renewable Energy* 177 (November): 967–75. <https://doi.org/10.1016/j.renene.2021.05.162>.
- Giudici, Mathyas, Luca Padalino, Giovanni Paolino, Ilaria Paratici, Alexandru Ionut Pascu, and Franca Garzotto. 2024. "Designing Home Automation Routines Using an LLM-Based Chatbot." *Designs* 8 (3): 3. <https://doi.org/10.3390/designs8030043>.

- Gomes, Isaías, Karol Bot, Maria Graça Ruano, and António Ruano. 2022. "Recent Techniques Used in Home Energy Management Systems: A Review." *Energies* 15 (8): 2866. <https://doi.org/10.3390/en15082866>.
- Gruver, Nate, Marc Finzi, Shikai Qiu, and Andrew G. Wilson. "Large language models are zero-shot time series forecasters." *Advances in Neural Information Processing Systems* 36 (2023): 19622-19635.
- Gupta, Anchal, Youakim Badr, Ashkan Negahban, and Robin G. Qiu. 2021. "Energy-Efficient Heating Control for Smart Buildings with Deep Reinforcement Learning." *Journal of Building Engineering* 34 (February): 101739. <https://doi.org/10.1016/j.jobbe.2020.101739>.
- Gutierrez-Martinez, Victor J., Carlos A. Moreno-Bautista, Jose M. Lozano-Garcia, Alejandro Pizano-Martinez, Enrique A. Zamora-Cardenas, and Miguel A. Gomez-Martinez. 2019. "A Heuristic Home Electric Energy Management System Considering Renewable Energy Availability." *Energies* 12 (4): 671. <https://doi.org/10.3390/en12040671>.
- Haase, Jennifer, and Paul H.P. Hanel. 2023. "Artificial Muses: Generative Artificial Intelligence Chatbots Have Risen to Human-Level Creativity." *Journal of Creativity* 33 (3): 100066. <https://doi.org/10.1016/j.yjoc.2023.100066>.
- Hadi, Muhammad Usman, Qasem Al Tashi, Rizwan Qureshi, et al. 2023. "A Survey on Large Language Models: Applications, Challenges, Limitations, and Practical Usage." Preprint, July 10. <https://doi.org/10.36227/techrxiv.23589741.v1>.
- Han, Binghui, Younes Zahraoui, Marizan Mubin, Saad Mekhilef, Mehdi Seyedmahmoudian, and Alex Stojcevski. 2023. "Home Energy Management Systems: A Review of the Concept, Architecture, and Scheduling Strategies." *IEEE Access* 11: 19999–20025. <https://doi.org/10.1109/ACCESS.2023.3248502>.
- Hasselt, Hado van, Arthur Guez, and David Silver. 2015. "Deep Reinforcement Learning with Double Q-Learning." arXiv:1509.06461. Preprint, arXiv, December 8. <http://arxiv.org/abs/1509.06461>.
- He, Jia, Mukund Rungta, David Koleczek, Arshdeep Sekhon, Franklin X. Wang, and Sadid Hasan. 2024. "Does Prompt Formatting Have Any Impact on LLM Performance?" arXiv:2411.10541. Preprint, arXiv, November 15. <https://doi.org/10.48550/arXiv.2411.10541>.
- Henderson, Peter, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. 2018. "Deep Reinforcement Learning That Matters." *Proceedings of the AAAI Conference on Artificial Intelligence* 32 (1). <https://doi.org/10.1609/aaai.v32i1.11694>.
- Huang, Chao, Hongcai Zhang, Long Wang, Xiong Luo, and Yonghua Song. 2022. "Mixed Deep Reinforcement Learning Considering Discrete-Continuous Hybrid Action Space for Smart Home Energy Management." *Journal of Modern Power Systems and Clean Energy* 10 (3): 743–54. <https://doi.org/10.35833/MPCE.2021.000394>.
- Jaihar, John, Neehal Lingayat, Patel Sapan Vijaybhai, Gautam Venkatesh, and K. P. Upla. 2020. "Smart Home Automation Using Machine Learning Algorithms." *2020 International Conference for Emerging Technology (INCET)*, June, 1–4. <https://doi.org/10.1109/INCET49848.2020.9154007>.
- Jelinek, Frederick. 1998. *Statistical Methods for Speech Recognition*. MIT Press.
- Ji, Ziwei, Tiezheng Yu, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. 2023. "Towards Mitigating LLM Hallucination via Self Reflection." In *Findings of the Association for Computational Linguistics: EMNLP 2023*, edited by Houda Bouamor, Juan Pino, and Kalika Bali. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.findings-emnlp.123>.

- Jiang, Xia, Jian Zhang, and Bo Wang. 2022. "Energy-Efficient Driving for Adaptive Traffic Signal Control Environment via Explainable Reinforcement Learning." *Applied Sciences* 12 (11): 5380. <https://doi.org/10.3390/app12115380>.
- Jin, Ming, Yifan Zhang, Wei Chen, Kexin Zhang, Yuxuan Liang, Bin Yang, Jindong Wang, Shirui Pan, and Qingsong Wen. 2024. "Position: What Can Large Language Models Tell Us about Time Series Analysis."
- Kakade, Sham, and John Langford. 2002. "Approximately Optimal Approximate Reinforcement Learning." *Proceedings of the Nineteenth International Conference on Machine Learning* (San Francisco, CA, USA), ICML '02, July 8, 267–74.
- Kaplan, Jared, Sam McCandlish, Tom Henighan, et al. 2020. "Scaling Laws for Neural Language Models." arXiv:2001.08361. Preprint, arXiv, January 22. <http://arxiv.org/abs/2001.08361>.
- Kaplar, Aleksandra, Milan Vidaković, Aleksandar Kaplar, Jovana Vidaković, and Jelena Slivka. 2024. "House Energy Management System, for Balancing Electricity Costs and Residential Comfort, Based on Deep Reinforcement Learning." *Acta Polytechnica Hungarica* 21 (8): 289–308. <https://doi.org/10.12700/APH.21.8.2024.8.15>.
- Khoury, Raphaël, Anderson R. Avila, Jacob Brunelle, and Baba Mamadou Camara. 2023. "How Secure Is Code Generated by ChatGPT?" arXiv:2304.09655. Preprint, arXiv, April 19. <https://doi.org/10.48550/arXiv.2304.09655>.
- King, Evan, Haoxiang Yu, Sangsu Lee, and Christine Julien. 2024. "Sasha: Creative Goal-Oriented Reasoning in Smart Homes with Large Language Models." *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 8 (1): 1–38. <https://doi.org/10.1145/3643505>.
- Kirk, James R., Robert E. Wray, Peter Lindes, and John E. Laird. 2024. "Improving Knowledge Extraction from LLMs for Task Learning through Agent Analysis." *Proceedings of the AAAI Conference on Artificial Intelligence* 38 (16): 18390–98. <https://doi.org/10.1609/aaai.v38i16.29799>.
- Kirschen, D.S., G. Strbac, P. Cumperayot, and D. de Paiva Mendes. 2000. "Factoring the Elasticity of Demand in Electricity Prices." *IEEE Transactions on Power Systems* 15 (2): 612–17. <https://doi.org/10.1109/59.867149>.
- Kontogiannis, Dimitrios, Dimitrios Bargiotas, and Aspasia Daskalopulu. 2021. "Fuzzy Control System for Smart Energy Management in Residential Buildings Based on Environmental Data." *Energies* 14 (3): 752. <https://doi.org/10.3390/en14030752>.
- Kostková, K., L. Omelina, P. Kyčina, and P. Jamrich. 2013. "An Introduction to Load Management." *Electric Power Systems Research* 95 (February): 184–91. <https://doi.org/10.1016/j.epsr.2012.09.006>.
- Koubaa, Anis. 2023. "GPT-4 vs. GPT-3.5: A Concise Showdown." Preprint, March 24. <https://doi.org/10.20944/preprints202303.0422.v1>.
- Kozlica, Reuf, Stefan Wegenkittl, and Simon Hirländer. 2023. "Deep Q-Learning versus Proximal Policy Optimization: Performance Comparison in a Material Sorting Task." arXiv:2306.01451. Preprint, arXiv, June 2. <https://doi.org/10.48550/arXiv.2306.01451>.
- Watkins, Christopher John Cornish Hellaby. "Learning from delayed rewards." (1989).
- Lee, Sangyoon, and Dae-Hyun Choi. 2019. "Reinforcement Learning-Based Energy Management of Smart Home with Rooftop Solar Photovoltaic System, Energy Storage System, and Home Appliances." *Sensors (Basel)* 19 (18): 3937. <https://doi.org/10.3390/s19183937>.

- Leitao, Joaquim, Paulo Gil, Bernardete Ribeiro, and Alberto Cardoso. 2020. "A Survey on Home Energy Management." *IEEE Access* 8: 5699–722. <https://doi.org/10.1109/ACCESS.2019.2963502>.
- Levy, Mosh, Alon Jacoby, and Yoav Goldberg. 2024. "Same Task, More Tokens: The Impact of Input Length on the Reasoning Performance of Large Language Models." arXiv:2402.14848. Preprint, arXiv, July 10. <https://doi.org/10.48550/arXiv.2402.14848>.
- Li, Hepeng, Zhiqiang Wan, and Haibo He. 2020. "A Deep Reinforcement Learning Based Approach for Home Energy Management System." *2020 IEEE Power & Energy Society Innovative Smart Grid Technologies Conference (ISGT)* (Washington, DC, USA), February, 1–5. <https://doi.org/10.1109/ISGT45199.2020.9087647>.
- Lin, Tianyang, Yuxin Wang, Xiangyang Liu, and Xipeng Qiu. 2022. "A Survey of Transformers." *AI Open* 3: 111–32. <https://doi.org/10.1016/j.aiopen.2022.10.001>.
- Lissa, Paulo, Conor Deane, Michael Schukat, Federico Seri, Marcus Keane, and Enda Barrett. 2021. "Deep Reinforcement Learning for Home Energy Management System Control." *Energy and AI* 3 (March): 100043. <https://doi.org/10.1016/j.egyai.2020.100043>.
- Liu, Pengfei, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. "Pre-Train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing." *ACM Computing Surveys* 55 (9): 1–35. <https://doi.org/10.1145/3560815>.
- Liu, Yuankun, Dongxia Zhang, and Hoay Beng Gooi. 2020. "Optimization Strategy Based on Deep Reinforcement Learning for Home Energy Management." *CSEE Journal of Power and Energy Systems* 6 (3): 572–82. <https://doi.org/10.17775/CSEEJPES.2019.02890>.
- Liu, Yuanyuan, Bo Qiu, Xiaodong Fan, Haijing Zhu, and Bochang Han. 2016. "Review of Smart Home Energy Management Systems." *Energy Procedia* 104 (December): 504–8. <https://doi.org/10.1016/j.egypro.2016.12.085>.
- Lonza, Andrea. 2019. *Reinforcement Learning Algorithms with Python: Learn, Understand, and Develop Smart Algorithms for Addressing AI Challenges*. Packt Publishing, Limited.
- Lundberg, Scott, and Su-In Lee. 2017. "A Unified Approach to Interpreting Model Predictions." arXiv:1705.07874. Preprint, arXiv, November 25. <http://arxiv.org/abs/1705.07874>.
- Mahapatra, Bandana, and Anand Nayyar. 2022. "Home Energy Management System (HEMS): Concept, Architecture, Infrastructure, Challenges and Energy Management Schemes." *Energy Syst* 13 (3): 643–69. <https://doi.org/10.1007/s12667-019-00364-w>.
- Maimó-Far, Aina, Alexis Tantet, Víctor Homar, and Philippe Drobinski. 2020. "Predictable and Unpredictable Climate Variability Impacts on Optimal Renewable Energy Mixes: The Example of Spain." *Energies* 13 (19): 5132. <https://doi.org/10.3390/en13195132>.
- Mbuwir, Brida V., Lennert Vanmunster, Klaas Thoelen, and Geert Deconinck. 2021. "A Hybrid Policy Gradient and Rule-Based Control Framework for Electric Vehicle Charging." *Energy and AI* 4 (June): 100059. <https://doi.org/10.1016/j.egyai.2021.100059>.
- Merrick, Luke, and Ankur Taly. 2020. "The Explanation Game: Explaining Machine Learning Models Using Shapley Values." arXiv:1909.08128. Preprint, arXiv, June 25. <https://doi.org/10.48550/arXiv.1909.08128>.

- Mersha, Melkamu, Khang Lam, Joseph Wood, Ali AlShami, and Jugal Kalita. 2024. "Explainable Artificial Intelligence: A Survey of Needs, Techniques, Applications, and Future Direction." *Neurocomputing* 599 (September): 128111. <https://doi.org/10.1016/j.neucom.2024.128111>.
- Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. "Efficient Estimation of Word Representations in Vector Space." arXiv:1301.3781. Preprint, arXiv, September 7. <https://doi.org/10.48550/arXiv.1301.3781>.
- Mirchandani, Suvir, Fei Xia, Pete Florence, et al. 2023. "Large Language Models as General Pattern Machines." arXiv:2307.04721. Preprint, arXiv, October 26. <https://doi.org/10.48550/arXiv.2307.04721>.
- Mishra, Swaroop, Daniel Khashabi, Chitta Baral, Yejin Choi, and Hannaneh Hajishirzi. 2022. "Reframing Instructional Prompts to GPTk's Language." arXiv:2109.07830. Preprint, arXiv, March 15. <https://doi.org/10.48550/arXiv.2109.07830>.
- Nikolić, Mladen, and Anđelka Zečević. "Mašinsko učenje." Beograd: Matematički fakultet (2019).
- Mnih, Volodymyr, Adrià Puigdomènech Badia, Mehdi Mirza, Alex Graves, Tim Harley, Timothy P Lillicrap, David Silver, and Koray Kavukcuoglu. 2016. "Asynchronous Methods for Deep Reinforcement Learning."
- Mnih, Volodymyr, Koray Kavukcuoglu, David Silver, et al. 2013. "Playing Atari with Deep Reinforcement Learning." arXiv, December. <https://doi.org/10.48550/arXiv.1312.5602>.
- Mnih, Volodymyr, Koray Kavukcuoglu, David Silver, et al. 2015. "Human-Level Control through Deep Reinforcement Learning." *Nature* 518 (7540): 529–33. <https://doi.org/10.1038/nature14236>.
- Van Moffaert, Kristof, and Ann Nowé. "Multi-objective reinforcement learning using sets of pareto dominating policies." *The Journal of Machine Learning Research* 15, no. 1 (2014): 3483-3512.
- Mohammadi, Mehdi, Ala Al-Fuqaha, Mohsen Guizani, and Jun-Seok Oh. 2018. "Semisupervised Deep Reinforcement Learning in Support of IoT and Smart City Services." *IEEE Internet of Things Journal* 5 (2): 624–35. <https://doi.org/10.1109/JIOT.2017.2712560>.
- Molnar, Christoph, Gunnar König, Julia Herbinger, et al. 2021. "General Pitfalls of Model-Agnostic Interpretation Methods for Machine Learning Models." arXiv:2007.04131. Preprint, arXiv, August 17. <https://doi.org/10.48550/arXiv.2007.04131>.
- Naveed, Humza, Asad Ullah Khan, Shi Qiu, et al. 2024. "A Comprehensive Overview of Large Language Models." arXiv:2307.06435. Preprint, arXiv, October 17. <http://arxiv.org/abs/2307.06435>.
- Thirunavukarasu, Arun James, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. "Large language models in medicine." *Nature medicine* 29, no. 8 (2023): 1930-1940
- OpenAI, Josh Achiam, Steven Adler, et al. 2024. "GPT-4 Technical Report." arXiv:2303.08774. Preprint, arXiv, March 4. <https://doi.org/10.48550/arXiv.2303.08774>.
- OpenAI. Accessed November 2, 2024. <https://platform.openai.com>.
- Oprea, Simona-Vasilica, and Adela Băra. 2024. "A Recommendation System for Prosumers Based on Large Language Models." *Sensors* 24 (11): 11. <https://doi.org/10.3390/s24113530>.
- Osborne, Martin J., and Ariel Rubinstein. 2006. *A Course in Game Theory*. 12. print. MIT Press.

- Pan, Tingzhe, Zean Zhu, Hongxuan Luo, et al. 2024. "Home Energy Management Strategy to Schedule Multiple Types of Loads and Energy Storage Device with Consideration of User Comfort: A Deep Reinforcement Learning Based Approach." *Frontiers in Thermal Engineering* 4 (June): 1391602. <https://doi.org/10.3389/fther.2024.1391602>.
- Pedrasa, Michael Angelo, Ted Spooner, and Iain MacGill. "An energy service decision-support tool for optimal energy services acquisition." Centre for Energy and Environmental Markets, UNSW, the university of new south wales (2010).
- Pipattanasomporn, Manisa, Murat Kuzlu, Saifur Rahman, and Yonael Teklu. 2014. "Load Profiles of Selected Major Household Appliances and Their Demand Response Opportunities." *IEEE Transactions on Smart Grid* 5 (2): 742–50. <https://doi.org/10.1109/TSG.2013.2268664>.
- Plappert, Matthias, Rein Houthoofd, Prafulla Dhariwal, et al. 2018. "Parameter Space Noise for Exploration." arXiv:1706.01905. Preprint, arXiv, January 31. <https://doi.org/10.48550/arXiv.1706.01905>.
- Raza, Ali, Li Jingzhao, Yazeed Ghadi, Muhammad Adnan, and Mansoor Ali. 2024. "Smart Home Energy Management Systems: Research Challenges and Survey." *Alexandria Engineering Journal* 92 (April): 117–70. <https://doi.org/10.1016/j.aej.2024.02.033>.
- Reif, Emily, Daphne Ippolito, Ann Yuan, Andy Coenen, Chris Callison-Burch, and Jason Wei. 2022. "A Recipe For Arbitrary Text Style Transfer with Large Language Models." arXiv:2109.03910. Preprint, arXiv, March 31. <http://arxiv.org/abs/2109.03910>.
- Renze, Matthew. 2024. "The Effect of Sampling Temperature on Problem Solving in Large Language Models." In *Findings of the Association for Computational Linguistics: EMNLP 2024*, edited by Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen. Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.findings-emnlp.432>.
- Rezaee Jordehi, Ahmad. 2020. "Enhanced Leader Particle Swarm Optimisation (ELPSO): A New Algorithm for Optimal Scheduling of Home Appliances in Demand Response Programs." *Artificial Intelligence Review* 53 (3): 2043–73. <https://doi.org/10.1007/s10462-019-09726-3>.
- Ritchie, Hannah, Pablo Rosado, and Max Roser. 2020. "Breakdown of Carbon Dioxide, Methane and Nitrous Oxide Emissions by Sector." *Our World in Data*, June 10. <https://ourworldindata.org/emissions-by-sector>.
- Sabbir M. 2024. "Comparative Analysis of Double Deep Q-Network (Double DQN) and Proximal Policy Optimization (PPO) for Lane-Keeping in Autonomous Driving." *International Journal of Science and Research Archive* 13 (2): 1207–22. <https://doi.org/10.30574/ijrsra.2024.13.2.2237>.
- Salih, Ahmed, Zahra Raisi-Estabragh, Ilaria Boscolo Galazzo, et al. 2025. "A Perspective on Explainable Artificial Intelligence Methods: SHAP and LIME." *Advanced Intelligent Systems* 7 (1): 2400304. <https://doi.org/10.1002/aisy.202400304>.
- Schulman, John, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. "Proximal Policy Optimization Algorithms." arXiv, August. <https://doi.org/10.48550/arXiv.1707.06347>.
- Sennrich, Rico, Barry Haddow, and Alexandra Birch. 2016. "Neural Machine Translation of Rare Words with Subword Units." *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1715–25. <https://doi.org/10.18653/v1/P16-1162>.
- Shakya, Ashish Kumar, Gopinatha Pillai, and Sohom Chakrabarty. 2023. "Reinforcement Learning Algorithms: A Brief Survey." *Expert Systems with*

- Applications* 231 (November): 120495.
<https://doi.org/10.1016/j.eswa.2023.120495>.
- Shanahan, Murray, Kyle McDonell, and Laria Reynolds. 2023. "Role-Play with Large Language Models." arXiv:2305.16367. Preprint, arXiv, May 25. <http://arxiv.org/abs/2305.16367>.
- Shi, Chufan, Haoran Yang, Deng Cai, et al. 2024. "A Thorough Examination of Decoding Methods in the Era of LLMs." arXiv:2402.06925. Preprint, arXiv, October 8. <https://doi.org/10.48550/arXiv.2402.06925>.
- Shuvo, Salman Sadiq, and Yasin Yilmaz. 2022. "Home Energy Recommendation System (HERS): A Deep Reinforcement Learning Method Based on Residents' Feedback and Activity." *IEEE Transactions on Smart Grid* 13 (4): 2812–21. <https://doi.org/10.1109/TSG.2022.3158814>.
- Siudek, Aleksandra, Anna M. Klepacka, Wojciech J. Florkowski, and Piotr Gradziuk. 2020. "Renewable Energy Utilization in Rural Residential Housing: Economic and Environmental Facets." *Energies* 13 (24): 6637. <https://doi.org/10.3390/en13246637>.
- Sun, Jianwen, Yan Zheng, Jianye Hao, Zhaopeng Meng, and Yang Liu. 2020. "Continuous Multiagent Control Using Collective Behavior Entropy for Large-Scale Home Energy Management." *AAAI* 34 (01): 922–29. <https://doi.org/10.1609/aaai.v34i01.5439>.
- Sun, Yinghan, Hongxi Wang, Hua Chen, and Wei Zhang. 2025. "A Practical Introduction to Deep Reinforcement Learning." arXiv:2505.08295. Preprint, arXiv, May 13. <https://doi.org/10.48550/arXiv.2505.08295>.
- Susnjak, Teo, and Timothy McIntosh. 2024. "ChatGPT: The End of Online Exam Integrity?" *Education Sciences* 14 (6): 656. <https://doi.org/10.3390/educsci14060656>.
- Sutton, Richard S., and Andrew G. Barto. Reinforcement Learning: An Introduction. 2nd ed. Cambridge, MA: The MIT Press, 2018.
- Thomas, George. "Introduction to the modbus protocol." *The Extension* 9, no. 4 (2008): 1-4.
- Typhoon HIL - Expert Hardware-in-the-Loop Solutions. Accessed September 3, 2024. <https://www.typhoon-hil.com/>.
- Vaswani, Ashish, Noam Shazeer, Niki Parmar, et al. 2023. "Attention Is All You Need." arXiv:1706.03762. Preprint, arXiv, August 2. <http://arxiv.org/abs/1706.03762>.
- Vojnović, Nikola, Vladan Krsman, Jovana Vidaković, et al. 2025. "Residential Load Flow Modeling and Simulation." *Applied System Innovation* 8 (5). <https://doi.org/10.3390/asi8050130>.
- Vojnović, Nikola, Jovana Vidaković, and Milan Vidaković. 2022. "Multi-Threaded Unbalanced Power Flow of Large-Scale Networks with Advanced Bus Classification." *Computers and Electrical Engineering* 101 (July): 108134. <https://doi.org/10.1016/j.compeleceng.2022.108134>.
- Vojnović, Nikola, Jovana Vidaković, and Milan Vidaković. 2023a. "Decoupled Load Flow for Large-Scale Multiphase Distribution Networks." *Sustainable Energy, Grids and Networks* 34 (June): 101049. <https://doi.org/10.1016/j.segan.2023.101049>.
- Vojnović, Nikola, Jovana Vidaković, and Milan Vidaković. 2023b. "Multi-Threaded Power Flow of Large-Scale Active Multiphase Distribution Networks." *Expert Systems with Applications* 227 (October): 120313. <https://doi.org/10.1016/j.eswa.2023.120313>.
- Wei, Jason, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-Thought Prompting Elicits

- Reasoning in Large Language Models. arXiv:2201.11903, 2022. <https://doi.org/10.48550/arXiv.2201.11903>
- Claude. Anthropic. Accessed April 20, 2024. <https://claude.ai/>.
- White, Jules, Quchen Fu, Sam Hays, et al. 2023. “A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT.” arXiv:2302.11382. Preprint, arXiv, February 21. <http://arxiv.org/abs/2302.11382>.
- Wu, Shijie, Ozan Irsoy, Steven Lu, et al. 2023. “BloombergGPT: A Large Language Model for Finance.” arXiv:2303.17564. Preprint, arXiv, December 21. <https://doi.org/10.48550/arXiv.2303.17564>.
- Xu, Jiangjiao, Hisham Mahmood, Hao Xiao, Enrico Anderlini, and Mohammad Abusara. 2021. “Electric Water Heaters Management via Reinforcement Learning With Time-Delay in Isolated Microgrids.” *IEEE Access* 9: 132569–79. <https://doi.org/10.1109/ACCESS.2021.3112817>.
- Xu, Xu, Youwei Jia, Yan Xu, Zhao Xu, Songjian Chai, and Chun Sing Lai. 2020. “A Multi-Agent Reinforcement Learning-Based Data-Driven Method for Home Energy Management.” *IEEE Transactions on Smart Grid* 11 (4): 3201–11. <https://doi.org/10.1109/TSG.2020.2971427>.
- Yang, Chengrun, Xuezhi Wang, Yifeng Lu, et al. 2024. “Large Language Models as Optimizers.” arXiv:2309.03409. Preprint, arXiv, April 15. <http://arxiv.org/abs/2309.03409>.
- Yang, Jingfeng, Hongye Jin, Ruixiang Tang, et al. 2024. “Harnessing the Power of LLMs in Practice: A Survey on ChatGPT and Beyond.” *ACM Transactions on Knowledge Discovery from Data* 18 (6): 1–32. <https://doi.org/10.1145/3649506>.
- Yu, Liang, Weiwei Xie, Di Xie, et al. 2020. “Deep Reinforcement Learning for Smart Home Energy Management.” *IEEE Internet of Things Journal* 7 (4): 2751–62. <https://doi.org/10.1109/JIOT.2019.2957289>.
- Zhang, Chiyuan, Oriol Vinyals, Remi Munos, and Samy Bengio. 2018. “A Study on Overfitting in Deep Reinforcement Learning.” arXiv:1804.06893. Preprint, arXiv, April 20. <https://doi.org/10.48550/arXiv.1804.06893>.
- Zhang, H. L., J. Baeyens, J. Degève, and G. Cacères. 2013. “Concentrated Solar Power Plants: Review and Design Methodology.” *Renewable and Sustainable Energy Reviews* 22 (June): 466–81. <https://doi.org/10.1016/j.rser.2013.01.032>.
- Zhao, Wayne Xin, Kun Zhou, Junyi Li, et al. 2024. “A Survey of Large Language Models.” arXiv:2303.18223. Preprint, arXiv, October 13. <http://arxiv.org/abs/2303.18223>.
- Zhao, Zhiheng, and Chanaka Keerthisinghe. 2020. “A Fast and Optimal Smart Home Energy Management System: State-Space Approximate Dynamic Programming.” *IEEE Access* 8: 184151–59. <https://doi.org/10.1109/ACCESS.2020.3023665>.
- Zheng, Zhuang, Zhankun Sun, Jia Pan, and Xiaowei Luo. 2021. “An Integrated Smart Home Energy Management Model Based on a Pyramid Taxonomy for Residential Houses with Photovoltaic-Battery Systems.” *Applied Energy* 298 (September): 117159. <https://doi.org/10.1016/j.apenergy.2021.117159>.
- Zhou, Bin, Wentao Li, Ka Wing Chan, et al. 2016. “Smart Home Energy Management Systems: Concept, Configurations, and Scheduling Strategies.” *Renewable and Sustainable Energy Reviews* 61 (August): 30–40. <https://doi.org/10.1016/j.rser.2016.03.047>.
- Zhou, Denny, Nathanael Schärli, Le Hou, et al. 2023. “Least-to-Most Prompting Enables Complex Reasoning in Large Language Models.” arXiv:2205.10625. Preprint, arXiv, April 16. <http://arxiv.org/abs/2205.10625>.
- Zhou, Xinlei, Han Du, Yongjun Sun, Haoshan Ren, Ping Cui, and Zhenjun Ma. 2023. “A New Framework Integrating Reinforcement Learning, a Rule-Based Expert

System, and Decision Tree Analysis to Improve Building Energy Flexibility.”
Journal of Building Engineering 71 (July): 106536.
<https://doi.org/10.1016/j.jobe.2023.106536>.

Zhou, Yongchao, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. “Large language models are human-level prompt engineers.” In *The Eleventh International Conference on Learning Representations*. 2022.

Biografija

Aleksandra Kaplar rođena je 1992. godine u Novom Sadu. Završila je gimnaziju “Jovan Jovanović Zmaj” u Novom Sadu 2011. godine. Četvorogodišnje akademske studije na Fakultetu tehničkih nauka u Novom Sadu upisala je 2011. godine na studijskom programu Računarstvo i automatika. Diplomirala je 2015. godine sa temom “Implementacija mobilnih agenata u agentskom okruženju”. Potom je upisala master studije 2015. godine na istom studijskom program. Master rad je odbranila 2016. sa temom “Simulacija saobraćaja upotrebom agentskih tehnologija u SARL programskom jeziku”. Od 2016. godine je student doktorskih studija na Fakultetu tehničkih nauka. Autor je i koautor više naučnih radova predstavljenih na domaćim i međunarodnim konferencijama.

Овај Образац чини саставни део докторске дисертације, односно докторског уметничког пројекта који се брани на Универзитету у Новом Саду. Попуњен Образац укоричити иза текста докторске дисертације, односно докторског уметничког пројекта.

План третмана података

Назив пројекта/истраживања
Аутоматски систем управљања уређајима паметне куће за балансирање трошкова електричне енергије и удобности становања заснован на методама дубоког учења
Назив институције/институција у оквиру којих се спроводи истраживање
Факултет техничких наука, Универзитет у Новом Саду
Назив програма у оквиру ког се реализује истраживање
Рачунарство и аутоматика – докторска дисертација
1. Опис података
1.1 Врста студије
Укратко описати тип студије у оквиру које се подаци прикупљају
Докторска дисертација
1.2 Врсте података
а) квантитативни
б) квалитативни
1.3. Начин прикупљања података
а) анкете, упитници, тестови
б) клиничке процене, медицински записи, електронски здравствени записи
в) генотипови: навести врсту _____
г) административни подаци: навести врсту _____
д) узорци ткива: навести врсту _____
ђ) снимци, фотографије: навести врсту _____
е) текст, навести врсту Актуелна литература у области истраживања

ж) мапа, навести врсту _____

з) остало: описати _____

1.3 Формат података, употребљене скале, количина података

1.3.1 Употребљени софтвер и формат датотеке:

а) Excel фајл, датотека _____

б) SPSS фајл, датотека _____

с) PDF фајл, датотека _____

д) Текст фајл, датотека _____

е) JPG фајл, датотека _____

ф) Остало, датотека _____

1.3.2. Број записа (код квантитативних података)

а) број варијабли _____

б) број мерења (испитаника, процена, снимака и сл.) _____

1.3.3. Поновљена мерења

а) да

б) не

Уколико је одговор да, одговорити на следећа питања:

а) временски размак измедју поновљених мера је _____

б) варијабле које се више пута мере односе се на _____

в) нове верзије фајлова који садрже поновљена мерења су именоване као _____

Напомене: _____

Да ли формати и софтвер омогућавају дељење и дугорочну валидност података?

а) Да

б) Не

Ако је одговор не, образложити _____

2. Прикупљање података

2.1 Методологија за прикупљање/генерисање података

2.1.1. У оквиру ког истраживачког нацрта су подаци прикупљени?

а) експеримент, навести тип _____

б) корелационо истраживање, навести тип _____

ц) анализа текста, навести тип **Анализа доступне литературе**

д) остало, навести шта _____

2.1.2 Навести врсте мерних инструмената или стандарде података специфичних за одређену научну дисциплину (ако постоје).

2.2 Квалитет података и стандарди

2.2.1. Третман недостајућих података

а) Да ли матрица садржи недостајуће податке? Да **Не**

Ако је одговор да, одговорити на следећа питања:

а) Колики је број недостајућих података? _____

б) Да ли се кориснику матрице препоручује замена недостајућих података? Да **Не**

в) Ако је одговор да, навести сугестије за третман замене недостајућих података

2.2.2. На који начин је контролисан квалитет података? Описати

Квалитет података је контролисан поређењем експерименталних и теоријских података

2.2.3. На који начин је извршена контрола уноса података у матрицу?

3. Третман података и пратећа документација

3.1. Третман и чување података

3.1.1. Подаци ће бити депоновани у _____ репозиторијум.

3.1.2. URL адреса _____

3.1.3. DOI _____

3.1.4. Да ли ће подаци бити у отвореном приступу?

а) Да

б) Да, али после ембарга који ће трајати до _____

в) Не

Ако је одговор не, навести разлог

3.1.5. Подаци неће бити депоновани у репозиторијум, али ће бити чувани.

Образложење

3.2. Метаподаци и документација података

3.2.1. Који стандард за метаподатке ће бити примењен?

3.2.1. Навести метаподатке на основу којих су подаци депоновани у репозиторијум.

Ако је потребно, навести методе које се користе за преузимање података, аналитичке и процедуралне информације, њихово кодирање, детаљне описе варијабли, записа итд.

3.3 Стратегија и стандарди за чување података

3.3.1. До ког периода ће подаци бити чувани у репозиторијуму?

3.3.2. Да ли ће подаци бити депоновани под шифром? Да Не

3.3.3. Да ли ће шифра бити доступна одређеном кругу истраживача? Да Не

3.3.4. Да ли се подаци морају уклонити из отвореног приступа после извесног времена?

Да Не

Образложити

4. Безбедност података и заштита поверљивих информација

Овај одељак МОРА бити попуњен ако ваши подаци укључују личне податке који се односе на учеснике у истраживању. За друга истраживања треба такође размотрити заштиту и сигурност података.

4.1 Формални стандарди за сигурност информација/података

Истраживачи који спроводе испитивања с људима морају да се придржавају Закона о заштити података о личности (https://www.paragraf.rs/propisi/zakon_o_zastiti_podataka_o_licnosti.html) и одговарајућег институционалног кодекса о академском интегритету.

4.1.2. Да ли је истраживање одобрено од стране етичке комисије? Да **Не**

Ако је одговор Да, навести датум и назив етичке комисије која је одобрила истраживање

4.1.2. Да ли подаци укључују личне податке учесника у истраживању? Да **Не**

Ако је одговор да, наведите на који начин сте осигурали поверљивост и сигурност информација везаних за испитанике:

- a) Подаци нису у отвореном приступу
 - б) Подаци су анонимизирани
 - ц) Остало, навести шта
-

5. Доступност података

5.1. Подаци ће бити

a) јавно доступни

б) доступни само уском кругу истраживача у одређеној научној области

ц) затворени

Ако су подаци доступни само уском кругу истраживача, навести под којим условима могу да их користе:

Ако су подаци доступни само уском кругу истраживача, навести на који начин могу приступити подацима:

5.4. Навести лиценцу под којом ће прикупљени подаци бити архивирани.

6. Улоге и одговорност

6.1. Навести име и презиме и мејл адресу власника (аутора) података

Александра Каплар, aleksandra.a@uns.ac.rs

6.2. Навести име и презиме и мејл адресу особе која одржава матрицу с подацима

6.3. Навести име и презиме и мејл адресу особе која омогућује приступ подацима другим истраживачима
