

О. К. ВИШНЕВСЬКИЙасистент кафедри програмного забезпечення
Національний університет «Львівська політехніка»
ORCID: 0009-0005-4857-9669**Л. М. ЖУРАВЧАК**доктор технічних наук, старший науковий співробітник,
професор кафедри програмного забезпечення
Національний університет «Львівська політехніка»
ORCID: 0000-0002-1444-5882

ПІДВИЩЕННЯ ЕНЕРГОЕФЕКТИВНОСТІ В РОЗУМНИХ БУДІВЛЯХ ЗА ДОПОМОГОЮ СТРАТЕГІЙ КОНТРОЛЮ З ВИКОРИСТАННЯМ ГЛИБОКОГО НАВЧАННЯ З ПІДКРІПЛЕННЯМ

У дослідженні розглянуто застосування стратегій керування для зниження енергоспоживання будівель. Згідно аналізу наявних публікацій більшість науковців застосовували або складні методи на основі моделей, або менш ефективні підходи з *Q-learning*. У роботі запропоновано новий підхід до регулювання системи опалення, вентиляції та кондиціювання повітря у адміністративних будівлях середнього розміру, що базується на інтелектуальному контролері з підкріпленням, розробленому на основі алгоритму проксимальної політики оптимізації політики (*Proximal Policy Optimization*), який не вимагає опису моделі. Методологія дослідження поєднує використання симуляцій у середовищі *EnergyPlus*, які дають змогу точно і динамічно відтворювати поведінку системи за різних сценаріїв управління. В дослідженні ми розглянули регулювання температури подачі повітря до системи. З метою удосконалення підходу керування спроектовано симуляційне середовище за допомогою бібліотеки *Gymnasium*, яка є ефективною платформою для реалізації та оптимізації алгоритмів навчання з підкріпленням (*Reinforcement learning*). На відміну від класичних методів оптимізації, що потребують точного математичного опису фізичних процесів, глибоке навчання з підкріпленням формує дії керування на основі спостереження залежностей між попередніми діями та їх впливом на стан системи. Ефективність розробленого контролера на основі глибокого навчання з підкріпленням порівняно з класичним методом керування на основі стратегії коригування температури подачі повітря відповідно до температури зовнішнього середовища. Отримані результати засвідчили суттєве зменшення енергоспоживання на 27,8 %, при забезпеченні належного мікроклімату в приміщеннях за такими показниками, як температура, вологість і концентрація CO₂. Запропонований підхід демонструє перспективу застосування методів навчання з підкріпленням для реалізації ефективних стратегій керування, що можуть бути впроваджені у реальних системах керування будівлями.

Ключові слова: навчання з підкріпленням, енергоефективність, навантаження, будівля, керування, агент, оптимізація.

О. К. VYSHNEVSKYYAssistant at the Department of Software Engineering
Lviv Polytechnic National University
ORCID: 0009-0005-4857-9669**L. M. ZHURAVCHAK**Doctor of Technical Sciences, Senior Research Fellow,
Professor at the Department of Software Engineering
Lviv Polytechnic National University
ORCID: 0000-0002-1444-5882

ADVANCED CONTROL STRATEGIES USING DEEP REINFORCEMENT LEARNING FOR IMPROVING ENERGY EFFICIENCY IN SMART BUILDINGS

This research investigates building control strategies for decreasing the energy consumption. A review of existing literature indicates that prior studies employed either complex, model-based approaches or relatively less efficient *Q-learning* techniques. To address these limitations, we propose a novel control method for Heating, Ventilation, and Air Conditioning (HVAC) systems in medium-sized office buildings, utilizing an intelligent reinforcement learning controller grounded in the model-free Proximal Policy Optimization algorithm (PPO). Our approach incorporates simulations conducted in the *EnergyPlus* environment, which allows for a precise and responsive depiction of HVAC system dynamics under various control conditions. A key aspect of the study is the regulation of supply air temperature.

To support the implementation and enhancement of our control strategy, we developed a co-simulation framework using the Gymnasium library. This environment served as a robust foundation for testing and fine-tuning reinforcement learning models. The performance of the proposed deep reinforcement learning-based controller was benchmarked against a conventional control strategy that adjusts supply air temperature based on outdoor air conditions. Experimental results showed a significant 27.8 % reduction in energy use, all while preserving indoor comfort in terms of temperature, humidity, and carbon dioxide levels. This work highlights the potential of reinforcement learning to simplify the deployment of advanced control techniques in real-world building energy management systems. Unlike traditional optimization methods that require detailed mathematical models of the system, deep reinforcement learning infers optimal control actions by analyzing the relationship between system states and the outcomes of selected actions.

Key words: reinforcement learning, energy efficiency, load, building, control, agent, optimization.

Постановка проблеми

Будівлі є одним із основних споживачів енергії та джерел викидів вуглекислого газу, на них припадає біля 30 % загального споживання порівняно з іншими секторами. Оскільки більшість наявних будівель залишатиметься в експлуатації і надалі, зменшення енергоспоживання у цих спорудах є важливим завданням для досягнення кліматичних цілей. Поточні стратегії передбачають значне підвищення темпів реновації для забезпечення енергоефективності, однак реалізація цих планів є складною задачею, що робить заплановані показники недостижними без впровадження нових підходів.

Системи опалення, вентиляції та кондиціонування повітря (Heat, Ventilation, and Air Conditioning, HVAC) відіграють ключову роль у забезпеченні комфортних умов у приміщеннях, водночас вони істотно впливають на загальне енергоспоживання будівлі. Проте завдяки ефективному управлінню енергоресурсами будівлі можуть не лише зменшити власне споживання, а й брати активну участь у програмах гнучкого енергоспоживання. Це передбачає перенесення використання енергії на періоди поза піковими навантаженнями, зниження максимального споживання та стабілізацію енергомережі.

Методи управління енергоспоживанням на основі моделей, зокрема модельно-прогнозне керування (Model Predictive Control, MPC), використовують математичні представлення теплових характеристик будівлі та її взаємодії з навколишнім середовищем. Однак такі підходи можуть спричиняти похибки та складнощі при застосуванні до різних будівель або в умовах, що відрізняються від початкових параметрів.

На відміну від цього методи безмодельного керування не потребують попередньої інформації про фізичні характеристики будівлі, а натомість формують свої стратегії шляхом безпосереднього навчання на основі отриманих даних та зворотного зв'язку. Одним із таких підходів є навчання з підкріпленням (Reinforcement Learning, RL), яке дає змогу створювати гнучкі та стійкі системи управління енергоспоживанням. У RL агент (контролер гнучкого керування) визначає оптимальну послідовність дій, аналізуючи попередній досвід методом проб і помилок, а не покладаючись на попередньо запрограмовані правила. Ця безмодельна властивість робить RL перспективним для розв'язання задач керування та оптимізації навіть у випадках, коли система недостатньо вивчена або доступна інформація є обмеженою. Глибоке навчання з підкріпленням (Deep Reinforcement Learning, DRL) поєднує методи глибокого навчання та RL. Використання глибоких нейронних мереж (Deep Neural Network, DNN) дає змогу агенту виявляти складні закономірності та приймати оптимальніші рішення.

Традиційні системи HVAC у будівлях зазвичай базуються на статичних алгоритмах керування, що використовують фіксовані порогові значення та прості евристичні правила. Такі підходи обмежують можливості оптимізації енергоспоживання. Натомість сучасні методи, зокрема, MPC та DRL, демонструють значний потенціал порівняно з традиційними правилами керування. Аналізуючи дані, DRL здатний автоматично знаходити приховані закономірності та оптимізувати політики управління. Дослідження показують, що використання як MPC, так і RL дає змогу суттєво знизити енергоспоживання будівель без погіршення комфорту для користувачів. Це підтверджує перспективність таких підходів у створенні більш енергоефективних та стійких будівельних систем.

Аналіз останніх досліджень і публікацій

Аналіз останніх досліджень і публікацій продемонстрував ефективність методів прогнозного керування для систем HVAC у зниженні енергоспоживання будівель та забезпеченні гнучкішої роботи інженерних систем. У дослідженні [1] запропоновано простий, доступний і масштабований підхід до реалізації MPC у комерційних будівлях зі застарілим обладнанням як альтернативу традиційним методам, базованим на метеоданих. Для зменшення кількості сенсорів і зниження навантаження на обробку даних система використовує температуру витяжного повітря для оцінки внутрішньої температури. Результати тестування підтвердили ефективність запропонованого підходу, за місяць навесні витрати на енергію знизили на 33 %. Оптимізація використання наявних ресурсів дала змогу покращити температурний контроль, скоротити витрати та інтегрувати будівлю в механізми гнучкого енергоспоживання без необхідності встановлення додаткових сенсорів.

Стратегії керування RL, що не вимагають наявності моделі, дають змогу агенту взаємодіяти з навколишнім середовищем у реальному часі і враховувати невизначеності. У роботі [2] проведено детальний аналіз застосування RL у сфері енергетики, де розглянуто різні підходи, їхні сильні та слабкі сторони. Було встановлено, що

хоча RL здатний підвищити ефективність енергосистем у середньому на 20 %, можливості цього підходу ще не повністю розкриті. Автори зауважили недостатнє використання сучасних методів DRL та надмірну залежність від базових алгоритмів, таких як Q-learning. Незважаючи на перспективність підходів, орієнтованих на дані, інтеграція RL у складні енергетичні системи залишається непростим завданням.

У дослідженні [3] запропоновано інноваційний підхід до управління енергоспоживанням у житлових будинках, що поєднує MPC та RL. Основна мета системи – використання теплових властивостей будівлі та акумуляторних батарей для зміщення енергоспоживання на періоди низьких тарифів і продажу надлишкової сонячної енергії назад у мережу. Для врахування похибок моделювання, невизначеностей метеопрогнозу та змінної поведінки користувачів розроблено параметризовану структуру MPC, яка наближено визначає оптимальну стратегію керування енергоспоживанням. Ця модель постійно вдосконалюється за допомогою модифікованого алгоритму детермінованого актора і критика із затримкою. Моделювання підтвердило, що запропонований метод ефективно балансує між комфортом користувачів і економічною вигодою навіть за умов неточних моделей та мінливих реальних умов.

Традиційні системи HVAC у будівлях здебільшого працюють за строго заданими правилами керування. У роботі [4] розглянуто потенціал навчання з підкріпленням як більш гнучкої та енергоефективної альтернативи. Запропоновано багатоагентну RL-архітектуру для оптимізації роботи HVAC-систем із урахуванням зворотного зв'язку від користувачів щодо теплового комфорту. Для підвищення ефективності навчання досліджено методи спільного використання параметрів між агентами та різні стратегії попереднього навчання. Це суттєво скоротило час, необхідний для набуття агентами ефективних стратегій керування. Отримані результати свідчать, що запропонований підхід дає змогу зменшити енергоспоживання на 6 % для всієї будівлі та до 8 % для окремих приміщень порівняно з традиційними алгоритмами керування. Водночас рівень комфорту для користувачів не лише зберігся, а подекуди навіть перевищив показники базової системи.

У дослідженні [5] запропоновано новий метод оптимізації енергоспоживання та теплового комфорту в будівлях із кількома зонами, що поєднує штучний інтелект та методи управління на основі правил для центрального повітрообробного пристрою. Для впровадження отриманої моделі в реальні системи автоматизації будівель розроблено методику вилучення чітких правил керування із процесу ухвалення рішень агентом. Ці правила пройшли тестування в середовищі моделювання та були порівняні з початковою RL-моделлю та традиційними методами управління відповідно до стандартів. Отримані результати показали, що контролер, створений на основі вилучених правил, досяг майже такого ж рівня енергоефективності, як і агент, базований на штучному інтелекті (ШІ), але при цьому був значно простішим у реалізації. Це підтверджує перспективність поєднання ШІ та керування на основі правил для впровадження передових стратегій управління енергією в реальних умовах.

Непередбачувана поведінка мешканців, зокрема, їхні шаблони використання гарячої води, є значним викликом для оптимізації енергоспоживання будівель. Традиційні системи керування зазвичай використовують консервативні стратегії, що приводить до надмірного споживання енергії. У дослідженні [6] запропоновано систему керування на основі навчання з підкріпленням, здатну адаптуватися до стохастичної поведінки користувачів. Метод базується на безмодельному RL-підході, що дає змогу переносити систему на різні будівлі без необхідності створення точних моделей кожної з них. Агент RL проходить попереднє навчання в режимі офлайн, використовуючи ймовірнісну модель споживання гарячої води для реалістичного відтворення поведінки користувачів та прискорення процесу навчання. Ефективність підходу підтверджено на основі реальних даних із житлового будинку, споживання енергії скоротили приблизно на 20 % порівняно з традиційними методами керування, при цьому зберегли рівень комфорту. Отримані результати підкреслюють потенціал RL-методів у системах управління енергоспоживанням, що враховують динамічну та нестабільну поведінку користувачів.

У роботі [7] досліджено потенціал механізмів гнучкого управління попитом у житлових будинках для зниження енергоспоживання. Автори підкреслюють необхідність повністю автоматизованих систем управління енергією та пропонують підхід, заснований на RL. Оптимізація роботи системи керування енергоспоживанням розглядається як RL-завдання, яке вирішується шляхом кластеризації побутових пристроїв і незалежної оптимізації графіків їхньої роботи. Порівняно з традиційними методами цей підхід має низку переваг: відсутність необхідності явного моделювання задоволеності користувачів, можливість проактивного планування завдань системою, гнучкість у формуванні запитів користувачами та зниження обчислювальної складності. Для перевірки ефективності підходу застосовано алгоритм Q-learning, результати якого продемонстрували перспективність запропонованої методики для розроблення інтелектуальних систем енергоменеджменту в житлових будинках.

У дослідженні [8] запропоновано метод оптимізації енергоменеджменту в розумних будівлях на основі DRL. Розроблено комплексну систему управління енергоспоживанням, яка враховує різні складові, зокрема, зберігання енергії, генерацію електроенергії сонячними панелями, заряджання електромобілів та роботу побутових приладів. Для реалізації стратегії оптимального розподілу ресурсів дослідники застосували Q-learning із урахуванням відповідних обмежень роботи системи. Результати моделювання продемонстрували ефективність підходу: запропонована стратегія DRL не лише задовольняє енергетичні потреби будівлі, а й забезпечує оптимальний

розподіл енергії. Запропонований метод перевершує традиційні алгоритми керування, підтримуючи рівень помилок у межах 10 %.

Традиційні системи управління будівлями стикаються з труднощами опрацювання зростаючих обсягів різнорідних даних, зокрема, від сенсорів та розкладів користувачів. У дослідженні [9] розглянуто проблему управління якістю повітря у приміщеннях за допомогою підходу, що поєднує DRL та семантичні знання. Запропоновано використання онтологічних графів знань для структурованого подання інформації про якість повітря в будівлях та взаємозв'язків між різними факторами, що на неї впливають. Такий підхід забезпечує контекстно-орієнтоване представлення стану будівлі. Отримані семантичні дані інтегруються у Deep Q-learning (DQL) модель, що дає змогу розробити інтелектуальну систему управління будівлею. Така система може ефективно контролювати якість повітря, адаптувати роботу повітрообробних агрегатів до змінних умов та забезпечувати баланс між комфортом мешканців і енергоефективністю. Автори наголошують, що цей підхід спрощує управління будівлею, зменшуючи складність прийняття рішень для операторів об'єкта.

У роботі [10] запропоновано систему управління енергоспоживанням на основі RL для розумних будівель, інтегрованих у «розумну» енергомережу. Така система враховує можливість обміну енергією з мережею, локальну генерацію відновлюваної енергії (наприклад, сонячні панелі), зберігання енергії та заряджання електромобілів із функцією *vehicle-to-grid*. Завдання оптимізації енергоспоживання сформульовано як процес ухвалення рішень Маркова, що включає множину станів, дій, ймовірності переходів та функцію винагороди. Алгоритм Q-learning дає змогу системі навчатися приймати оптимальні рішення щодо розподілу енергії в умовах невизначеності, зокрема, з урахуванням змін попиту, заряджання електромобілів та генерації сонячної енергії. Моделювання на основі реальних даних показало, що запропонований метод знижує витрати на енергію порівняно з традиційними підходами та випадковими рішеннями. Автори зазначають, що ця RL-методика може бути адаптована для різних середовищ, зокрема, мікромереж та промислових об'єктів.

У дослідженні [11] запропоновано багатоагентну систему управління, спрямовану на оптимізацію комфорту в будівлях та підвищення енергоефективності. Її структура включає три незалежні агенти, кожен із яких відповідає за певний аспект внутрішнього середовища: якість повітря, візуальний та тепловий комфорт. Для оцінки присутності мешканців використано стохастичну модель, що застосовує ймовірнісні та еволюційні алгоритми, а CO₂-датчики виступають основним джерелом даних. Нечітка логіка використовується для врахування невизначеності параметрів системи. Контрольні агенти застосовують Fuzzy Q-learning для роботи з безперервними змінними стану та керування. Моделювання підтвердило ефективність підходу, продемонструвавши точне оцінювання присутності мешканців та скорочення енергоспоживання до 56 % без суттєвого зниження рівня комфорту.

Метод DQL використовує штучні нейронні мережі замість традиційної Q-таблиці, що дає змогу працювати з більшими просторами станів та дій. У роботі [12] запропоновано систему на основі нейронної Q-learning моделі, яка здатна адаптивно вивчати закономірності споживання енергії в домогосподарствах, зменшуючи пікове навантаження та сприяючи енергозбереженню. В основі методу лежить Neural Fitted Q-learning, що забезпечує швидке й ефективне прийняття рішень щодо енергоспоживання, зберігаючи баланс між мінімізацією витрат та комфортом користувачів. Моделювання на основі класичного канадського будинку показало значне скорочення енергоспоживання в пікові періоди, що сприяє зниженню викидів та екологічному сталому розвитку. Автори наголошують, що масштабне впровадження такого підходу може суттєво знизити піковий попит і сприяти ефективнішому розподілу ресурсів.

Дослідження [13] присвячено оптимізації розкладу енергоспоживання в житлових будівлях за допомогою DRL. Використано алгоритми DQL та Deep Policy Gradient для управління електроспоживанням, зокрема, в умовах змінних тарифів на електроенергію. Аналіз датасету Pecan Street підтвердив ефективність запропонованого підходу, алгоритм Deep Policy Gradient доцільніший за DQL для задач реального часу, оскільки забезпечує мінімізацію витрат та згладжування навантаження.

Застосування RL у реальних будівлях супроводжується викликами, зокрема, тривалими фазами навчання та високими операційними витратами. У дослідженні [14] розглянуто використання Transfer Learning (TL) для адаптації RL-контролерів до нових середовищ. Запропоновано онлайн TL-метод, що поєднує імітаційне навчання та донавчання моделі для перенесення контролера між двома офісними будівлями. Попередньо натренований у цифровому двійнику контролер продемонстрував енергозбереження на рівні 6–40 % та покращення температурного контролю на 30–50 %.

Згідно аналізу наукових джерел більшість авторів використовували або менш результативне Q-learning, або складні методи, засновані на математичних моделях. У запропонованій роботі реалізовано прогресивний підхід до керування на основі навчання з підкріпленням, який базується на сучасному алгоритмі проксимальної оптимізації політики (*Proximal Policy Optimization*, PPO), що дає змогу підвищити енергоефективність будівлі, зберігаючи вимоги до комфорту мешканців. Отже, результати цього дослідження сприятимуть удосконаленню підходів до розроблення надійних енергоефективних систем керування.

Формулювання мети дослідження

Метою цього дослідження є розроблення удосконаленої стратегії керування HVAC системою будівлі на основі безмодельного алгоритму навчання з підкріпленням для підвищення енергоефективності порівняно з традиційними підходами.

Завдання дослідження: проаналізувати сучасні підходи до керування HVAC системами, розробити архітектуру для середовища ко-симуляції з навчанням з підкріпленням та реалізувати його інтеграцію з будівельною симуляцією EnergyPlus, реалізувати базову стратегію керування на основі традиційного методу коригування температури зовнішнього повітря, розробити контролер керування HVAC системою будівлі на основі глибокого навчання з підкріпленням, що ґрунтується на безмодельному алгоритмі PPO, порівняти ефективність базової стратегії з результатами, отриманими за допомогою запропонованого інтелектуального контролера.

Викладення основного матеріалу дослідження

Методологія

У цьому дослідженні застосовано сучасний безмодельний підхід із використанням навчання з підкріпленням [15]. RL – це тип машинного навчання, за якого агент (контролер) навчається знаходити найкращий спосіб розв’язання задачі шляхом експериментів [16]. Контролер навчається, отримуючи винагороду від своїх дій, позитивну або негативну, залежно від ефекту на середовище. Процес навчання агента полягає у визначенні найкращої дії для кожного можливого стану з метою максимізації загальної винагороди [17]. Метою агента є розроблення стратегії, яка забезпечує найвищу сумарну винагороду з часом. Це передбачає певний компроміс, агент має збалансувати дослідження нових, потенційно кращих стратегій із використанням уже набутих знань [18].

Проксимальна оптимізація політики

У цьому дослідженні використано алгоритм PPO, який є гнучким і поширеним методом RL для розв’язання задач керування. З 2018 року PPO є алгоритмом навчання з підкріпленням в OpenAI за замовчуванням. Його ключова перевага полягає в ефективному балансуванні між дослідженням нових стратегій та використанням наявних знань, що є критично важливим для оптимізації продуктивності в складних середовищах. PPO застосовують у різних галузях, зокрема, в управлінні енергетичними системами, оскільки він демонструє кращу відповідність гіперпараметрам і вищу ефективність тренування порівняно з алгоритмом DDPG (Deep Deterministic Policy Gradient). Крім того, він забезпечує вигідніші значення цільової функції. Дослідження показали, що алгоритм PPO ефективно працює з неперервними просторами дій і має стабільніший механізм оновлення, що дає змогу знизити експлуатаційні витрати порівняно з DDPG. Зокрема, при застосуванні до задач енергетичного планування алгоритм PPO сприяв зменшенню енергетичних витрат порівняно з методами управління на основі MPC [19].

Алгоритм PPO підвищує стабільність навчання агента RL шляхом запобігання надмірним змінам політики. Він використовує коефіцієнт, що відображає розбіжність між поточною та попередньою політиками, та обмежує його в межах визначеного діапазону. Такий механізм “відсікання” гарантує, що оновлення політики залишаються помірними, що сприяє більшій стабільності в процесі навчання. Основна ідея PPO полягає у покращенні стабільності тренування шляхом обмеження величини змін політики на кожному етапі навчання, запобігаючи різким та потенційно негативним змінам. Це досягається за допомогою цільової функції, яка обмежує зміни політики у вузькому діапазоні. Такий підхід запобігає значним оновленням ваг, що могли б дестабілізувати процес навчання.

Марковський процес прийняття рішень

Задачу навчання з підкріпленням можна сформулювати як Марковський процес прийняття рішень (Markov Decision Process, MDP) [20, 21]. Для нашої задачі MDP можна визначити такими елементами:

- $\{S\}$ – множина можливих станів середовища (простір станів);
- $\{A\}$ – множина можливих дій агента (простір дій);
- $P_a(s_t, s_{t+1})$ – ймовірність переходу зі стану s_t у стан s_{t+1} після виконання дії a ;
- $R_a(s_t, s_{t+1})$ – винагорода, отримана агентом після переходу зі стану s_t у стан s_{t+1} внаслідок виконання дії a .

На кожному кроці часу t агент, слідуючи певній політиці π , взаємодіє із середовищем, виконуючи дію a_t , визначену відповідно до стану s_t спостереження:

$$a_t = \pi(s_t).$$

Середовище реагує на цю дію, повертаючи винагороду R_{a_t} , і переходить до стану s_{t+1} . Метою агента є навчитися оптимальної політики π , яка максимізує сукупну винагороду протягом усієї послідовності взаємодій:

$$\max E \left[\sum_{t=0}^{\infty} \gamma_t R_{a_t}(s_t, s_{t+1}) \right],$$

де γ – коефіцієнт дисконтування, що визначає вагу майбутніх винагород порівняно з попередніми ($0 \leq \gamma \leq 1$).

Простір станів

На кожному кроці часу агент отримує спостереження від середовища, які відображають поточний стан системи в цей момент часу, визначений так:

$$s_t = \{T_{out}, T_{in}, H_{in}, C_{in}, T_{hig_spt}, T_{clg_spt}, E_{hig_load}, E_{clg_load}\}. \quad (1)$$

Процес прийняття рішень глибоким агентом з підкріпленням враховує вісім різних змінних, вимірних в експериментальній будівлі під час симуляції з кроком часу $t = 15$ хвилин. Ці змінні разом з їхніми дозволеними діапазонами та одиницями вимірювання наведено в табл. 1.

Таблиця 1

Змінні простору станів

Змінні	Мін.	Макс.	Одиниця	Опис
T_{out}	-40	40	°C	Температура зовнішнього повітря
T_{in}	0	40	°C	Температура повітря приміщення
H_{in}	0	100	%	Відносна вологість повітря приміщення
C_{in}	0	100 000	ppm	Концентрація CO ₂ в повітрі
T_{hig_spt}	0	30	°C	Температура заданого значення термостата опалення
T_{clg_spt}	0	30	°C	Температура заданого значення термостата охолодження
E_{hig_load}	0	28	кВт·год	Навантаження на опалення (природний газ)
E_{clg_load}	0	28	кВт·год	Навантаження на охолодження (електроенергія)

Простір дій

Агент виконує на кожному кроці часу t (кожні 15 хвилин) дії, що представляють рішення щодо керування системою будівлі з використанням пристроїв для опалення та охолодження. У цьому дослідженні агент визначає оптимальну температуру подачі повітря як контрольну змінну, яка служить механізмами агента для оптимізації енергоефективності. Таким чином, дії на кроці часу t можуть бути представлені рівнянням:

$$a_t = \{T_{sat}\}. \quad (2)$$

Простір дій дискретний і складається з 100 можливих варіантів. У нашому дослідженні ці дискретні дії відображають неперервний діапазон від 15.0 до 30.0, що відповідає діапазону заданих температур подавання повітря.

Перехід між станами

Коли агент приймає рішення щодо керування, це впливає на стан середовища, що може призвести до нового стану. Перехід до конкретного нового стану залежить не лише від виконаної дії, але й від різних непередбачуваних факторів у середовищі. Моделювання цих переходів може бути дуже складним через невизначеності у середовищі [22]. Однак DRL пропонує спосіб уникнути цієї проблеми шляхом навчання безпосередньо з одержаного досвіду. Замість того, щоб намагатися явно моделювати ймовірність кожного переходу, DRL використовує нейронні мережі для визначення ефектів невизначеності на основі даних спостережень [23].

Функція винагороди

Основною функцією DRL контролера є оптимізація температури подавання повітря (T_{sat}) шляхом регулювання цієї змінної кожні 15 хвилин. Температура подавання повітря встановлюється в межах від 15 °C до 30 °C. Процес оптимізації має на меті зменшити дискомфорт температури в зоні, а також підтримувати вологість у межах прийнятних значень, мінімізуючи споживання енергії та рівень CO₂. Функція винагороди R_t , яка використовується для DRL агента, визначена так:

$$R_t = -(E_{penalty} + Temp_{penalty} + Hum_{penalty} + CO_{2_penalty}), \quad (3)$$

де:

$$\begin{aligned}
 R_{penalty} &= -(E_{heating_load} + E_{cooling_load}), \\
 Temp_{penalty} &= \begin{cases} \min((T_{in} - T_{clg_spt}), (T_{hig_spt} - T_{in})), & \text{if } (T_{in} < T_{clg_spt} \text{ or } T_{in} > T_{hig_spt}) \\ 0 & \text{else} \end{cases}, \\
 Hum_{penalty} &= \begin{cases} \min((H_{in} - 40), (60 - H_{in})), & \text{if } (H_{in} < 40 \text{ or } H_{in} > 60) \\ 0 & \text{else} \end{cases}, \\
 CO_{2_penalty} &= C_{in}.
 \end{aligned}$$

У такому формулюванні функції винагороди процес прийняття рішень у рамках MDP спрямований на визначення розкладів енергетичного управління, які максимізують кумулятивну винагороду протягом часу, в результаті чого зменшуються експлуатаційні витрати [24].

Енергетичне споживання відображає загальне споживання енергії системою HVAC будівлі, що поєднує навантаження на опалення (природний газ) та охолодження (електрика). Дискомфорт температури в зоні мінімізується шляхом накладення штрафу (зменшення винагороди) на відхилення температури повітря в приміщенні T_{in} від встановленої

температури для опалення T_{hig_spt} та охолодження T_{clg_spt} . Коли температура в кімнаті опускається нижче температури опалення, штраф збільшується пропорційно тому, наскільки вона занадто низька. Однак, коли температура перевищує встановлену температуру для охолодження, штраф збільшується пропорційно, відображаючи дискомфорт від занадто високої температури, особливо в літні періоди, коли охолодження необхідне. Для оптимального рівня вологості накладається штраф, коли вологість H_m виходить за межі визначеного комфортного діапазону від 40 % до 60 %. Цей штраф збільшується відповідно до того, наскільки вологість відхиляється від дозволеного діапазону.

Розробка контролера з підкріпленням

У цьому підрозділі описано процес розроблення контролера з підкріпленням, який приймає керуючі рішення на основі попередньо навченої оптимальної політики та передає сигнали управління до змодельованої системи HVAC. Контролер реалізовано мовою програмування Python 3.11 з використанням фреймворку Gymnasium 0.28.1, що підтримується Farama Foundation і є форком OpenAI Gym [25]. Gymnasium спеціально розроблено для створення та тестування алгоритмів навчання з підкріпленням у конфігурованому просторі, де агенти взаємодіють з навколишнім середовищем за допомогою механізмів дій і винагород, навчаючись та коригуючи свою політику керування.

У межах дослідження розроблено кастомізоване середовище Gymnasium для симуляційної платформи EnergyPlus 24.2 – комплексного засобу енергетичного моделювання будівель, що дає змогу оцінювати споживання енергії з урахуванням потреб у опаленні, охолодженні та вентиляції.

Простори станів та дій сформовано відповідно до (1) та (2). Функцію винагороди визначено згідно з (3), що дає змогу кількісно оцінити ефективність дій агента для досягнення цільових показників – мінімізації енергоспоживання за умови збереження комфортних умов. Функція винагороди є ключовим елементом, що спрямовує процес навчання агента.

Схематична структура створеного середовища коесимуляції представлена на рис. 1.

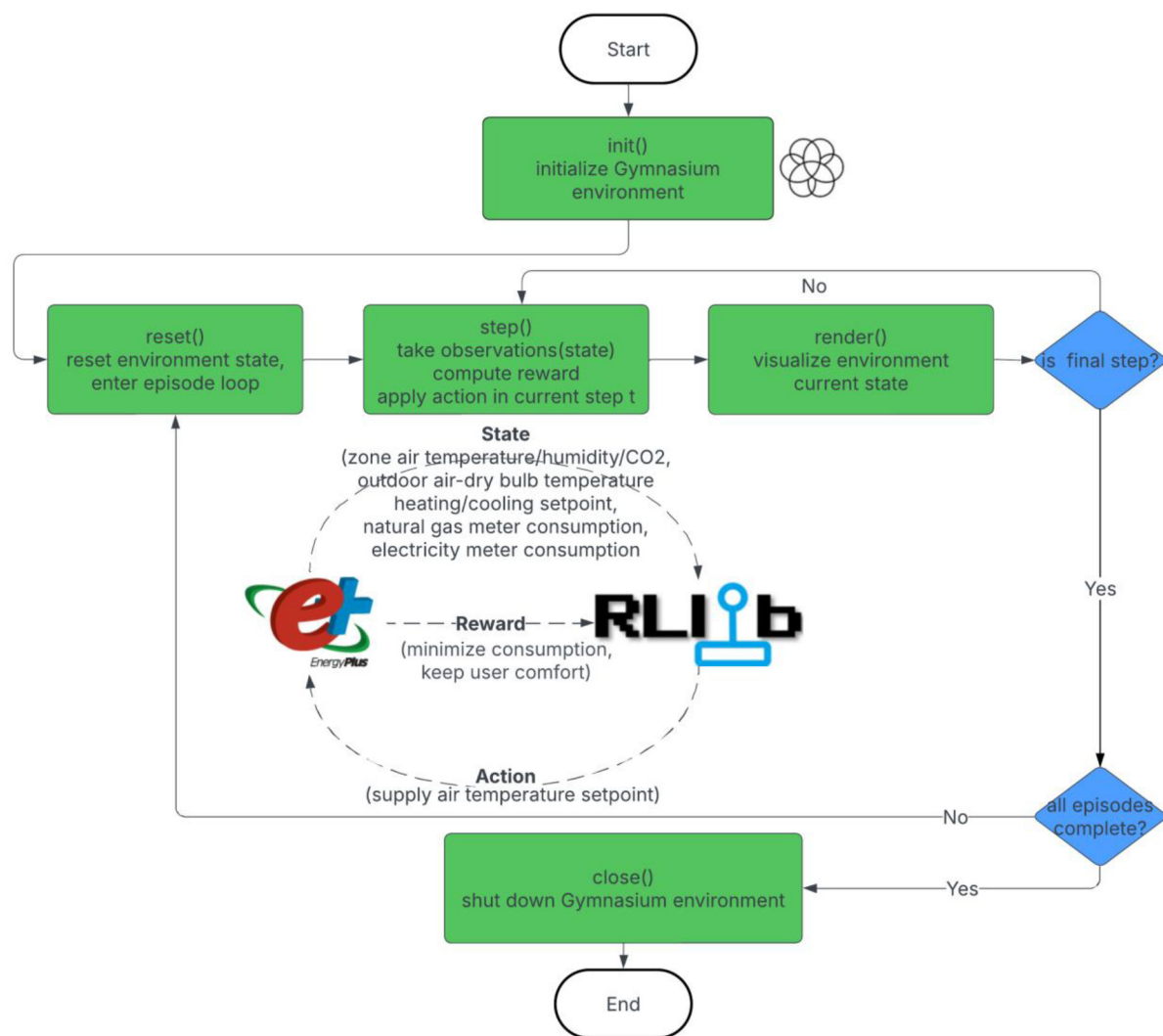


Рис. 1. Схеми навчання з підкріпленням у поєднанні з симуляцією EnergyPlus

Для забезпечення програмного доступу до процесу симуляції в EnergyPlus ми використали EnergyPlus Python API, який надає інтеграційні можливості з фреймворком Gymnasium. Це дає змогу середовищу Gymnasium надсилати команди до симуляції EnergyPlus та отримувати зворотні дані. Така інтеграція формує двонаправлений канал зв'язку, де змінні стани симуляції постійно передаються до алгоритму навчання, а дії, запропоновані агентом, реалізуються в середовищі EnergyPlus.

Для налаштування алгоритму в межах цього спеціалізованого середовища використано бібліотеку Ray Reinforcement Learning Library (RLib) версії 2.20.2. Rlib – це високорівнева відкрита бібліотека, призначена для масштабованих, готових до впровадження алгоритмів навчання з підкріпленням, зокрема, алгоритму PPO та інших [26].

У ході дослідження застосовано такі гіперпараметри для налаштування алгоритму PPO: коефіцієнт дисконтування $\gamma = 0.95$, швидкість навчання $lr = 0.003$, оптимізатор – Adam, початковий коефіцієнт KL-дивергенції $kl_coeff = 0.3$, розмір навчального пакету $train_batch_size = 2880$, розмір мініпакету стохастичного градієнтного спуску $sgd_minibatch_size = 360$, коефіцієнт втрат функції цінності $vf_loss_coeff = 0.01$, параметр обрізання політики $clip_param = 0.2$, використання критика як базової функції $use_critic = true$, застосування генералізованого оцінювання переваги (GAE) $use_gae = true$. Також активовано шар довготривалої короткочасної пам'яті (LSTM) для покращення продуктивності в задачах із часовою залежністю. Для навчання використовувався фреймворк Torch.

Проведено низку циклів навчання та валідації з метою вдосконалення моделі. Ці цикли включали ітераційні запуски симуляцій, під час яких політика агента постійно адаптувалась на основі зворотного зв'язку з середовищем. Процес навчання тривав до досягнення певної кількості кроків $timesteps = 2\,880\,000$ (1000 епізодів), після чого зберігалася найкраща політика.

Розроблений додаток для косимуляції має консольний інтерфейс. Приклад роботи, вивід поточної ітерації, епізоду та значенням винагороди наведено на рис. 2. EnergyPlus дає опцію генерації вихідних файлів у форматі «csv», що містять результати симуляції. Зокрема, після завершення симуляції формується файл «eplusout.csv», який містить інформацію про значення змінних, встановлені параметри та лічильники на всіх кроках моделювання. Ці дані використано для побудови графіків та візуалізації процесу симуляції. Бібліотека Ray здійснює логування результатів у TensorBoard, що дає змогу аналізувати метрики навчання на всіх ітераціях.

```

Select Administrator: Node.js command prompt - poetry run rllib --env OfficeEnv
(RolloutWorker pid=83116) Calculate Outside Surface Heat Balance [repeated 6x across cluster]
(RolloutWorker pid=83116) Calculate Inside Surface Heat Balance [repeated 6x across cluster]
(RolloutWorker pid=83116) Calculate Air Heat Balance [repeated 6x across cluster]
(RolloutWorker pid=83116) Initializing HVAC [repeated 6x across cluster]
(RolloutWorker pid=84024) EnergyPlus Completed Successfully. [repeated 6x across cluster]
Trial status: 1 RUNNING
(RolloutWorker pid=83116) Warming up {3} [repeated 20x across cluster]
Current time: 2025-04-22 23:07:25. Total running time: 1min 30s
(RolloutWorker pid=83116) Calculating Detailed Daylighting Factors, Start Date=01/01 [repeated 6x across cluster]
Logical resource usage: 3.0/16 CPUs, 0/1 GPUs (0.0/1.0 accelerator_type(s))
(RolloutWorker pid=84024) Performing Zone Sizing Simulation [repeated 4x across cluster]

+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
| Trial name           | status  | iter  | total time (s) | ts      | reward  | episode_reward_max | episode_reward_min | episode_len_mean | episodes_this_iter |
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+
| PPO_OfficeEnv_3286f_00000 | RUNNING | 34    | 68.2288       | 19584   | 70.7637 | 100.197            | 46.5913            | 288              | 2                  |
+-----+-----+-----+-----+-----+-----+-----+-----+-----+-----+

(RolloutWorker pid=83116) ...for Sizing Period: #2 SIZING PERIOD DESIGN DAY 2 [repeated 6x across cluster]
(RolloutWorker pid=83116) Calculating System sizing [repeated 4x across cluster]
(RolloutWorker pid=84024) Adjusting Air System Sizing [repeated 2x across cluster]
(RolloutWorker pid=84024) Adjusting Standard 62.1 Ventilation Sizing [repeated 2x across cluster]
(RolloutWorker pid=84024) Initializing Simulation [repeated 2x across cluster]
(RolloutWorker pid=83116) Reporting Surfaces [repeated 2x across cluster]
(RolloutWorker pid=83116) Calculating Detailed Daylighting Factors [repeated 2x across cluster]
(RolloutWorker pid=84024) Initializing New Environment Parameters [repeated 2x across cluster]
(RolloutWorker pid=83116) Calculating Daylighting Coefficients (Map Points), Zone-THERMAL_ZONE 1 [repeated 2x across cluster]
(RolloutWorker pid=83116) Starting Simulation at 01/01/2023 for RUN PERIOD 1 [repeated 3x across cluster]
(RolloutWorker pid=83116) Writing tabular output file results using comma format. [repeated 4x across cluster]
(RolloutWorker pid=83116) Writing tabular output file results using HTML format. [repeated 4x across cluster]
(RolloutWorker pid=83116) Computing Life Cycle Costs and Reporting [repeated 4x across cluster]
(RolloutWorker pid=83116) Writing final SQL reports [repeated 4x across cluster]
(RolloutWorker pid=84024) ReadVarsESO program starting. [repeated 6x across cluster]
(RolloutWorker pid=84024) Requested ESO file=c:\Users\OVYSHN-1\AppData\Local\Temp\tmpz7stwp5j\episode-00000126-77184\eplusout.mtr [repeated 6x across cluster]

```

Рис. 2. Процес виконання розробленого додатку косимуляції

Експериментальні дослідження

Експерименти реалізовано на ноутбуці Asus Vivobook Pro, оснащеному 8-ядерним процесором AMD Ryzen 7 6800H з тактовою частотою 3.20 ГГц, 32 ГБ оперативної пам'яті та 12-ядерним графічним процесором AMD Radeon 680M під керуванням операційної системи Windows 11 Pro.

У цьому дослідженні використано середовище симуляції будівлі на основі моделювання в EnergyPlus версії 24.2.0. Для симуляції застосовано модель середнього офісного будинку у форматі «idf» з прикладів OpenStudio Application версії 1.8.0, що продемонстровано на рис. 2. Дані про погодні умови для симуляції використано для львівського міжнародного аеропорту за 2009–2023 роки (файл «UKR_LV_Lviv.Intl.AP.333930_TMYx.2009-2023.erw») із репозиторію кліматичних даних для симуляцій будівель. Будівля має один поверх висотою 3 метри, чотири приміщення шириною і довжиною по 10 м, зовнішні стіни, зведені із цегли товщиною 100 мм, дах з легкого бетону товщиною 100 мм, два вікна і одні металеві двері (рис. 3).

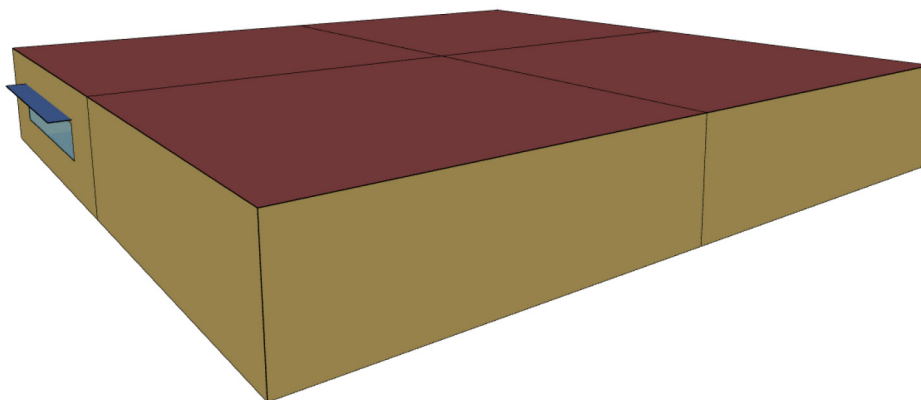


Рис. 3. Модель офісної будівлі, використана для симуляції

Модель будівлі оснащена системою HVAC з повітряним контуром, яка обслуговує теплову зону (рис. 4).

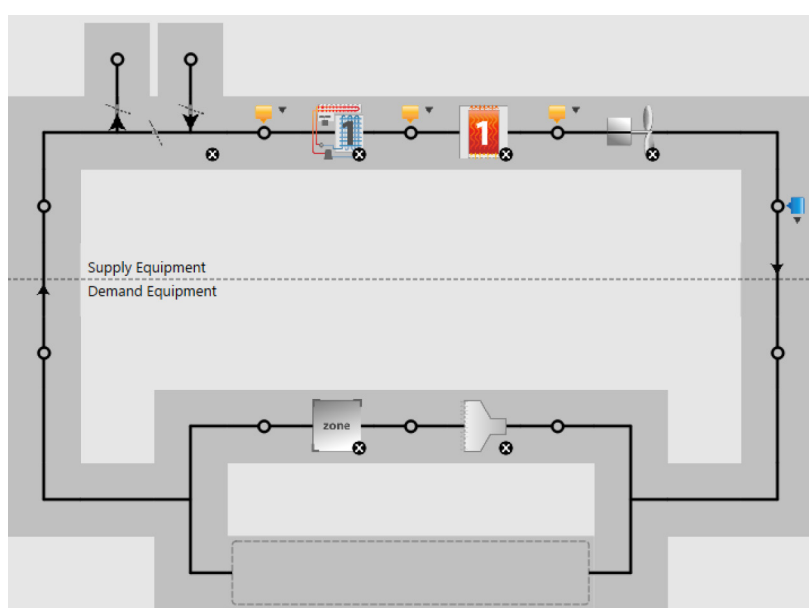


Рис. 4. Система HVAC з повітряним контуром

Система включає базовий охолоджувальний блок із прямим розширенням з одношвидкісним компресором. Нагрівання в повітряному контурі забезпечується газовим нагрівальним елементом. Рециркуляція повітря здійснюється за допомогою вентилятора. Система припливного повітря відповідає за подачу зовнішнього повітря та вентиляцію основного повітряного контуру. Вона забезпечує змішування свіжого зовнішнього повітря з повітрям, що повертається з приміщень, після чого змішане повітря піддається кондиціонуванню (охолодженню або нагріванню) перед повторною подачею в зони. Система працює з постійною витратою повітря без можливості регулювання подавання або повторного нагріву. Модель включає чотири приміщення з визначеними тепловими навантаженнями від людей, освітлення та електричного обладнання. Розклади для освітлення, кількості людей та електричних приладів, а також охолоджувальних і нагрівальних уставок змінюються залежно від часу доби, робочих днів та святкових періодів.

Симуляція в EnergyPlus охоплює один календарний місяць, з 1 по 30 квітня 2023 року. Для моделювання встановлено крок симуляції $timestep = 4$, що відповідає інтервалу 15 хвилин. Інші параметри моделі будівлі та налаштування симуляції наведено в табл. 2.

Проведено серію експериментів з налаштуванням гіперпараметрів агента глибокого навчання в рамках розробленого середовища ко-симуляції, починаючи з короткотривалого симуляційного періоду (1 день), 96 000 кроків часу та 1000 епізодів. Це дало змогу алгоритму РРО частіше оновлювати політику, що спрощує агенту навчання щоденних патернів, зменшує варіативність досвіду та забезпечує стабільніший навчальний сигнал завдяки налаштованій функції винагороди. Ми застосовували підхід поступового навчання, збільшуючи тривалість симуляції, коли ефективність агента покращується. Це пришвидшило цикл розробки та тестування, при цьому розширення

Таблиця 2

Параметри симуляції моделі будівлі

Параметр	Значення	Одиниця
Кількість людей	0,05	людей/м ²
Швидкість утворення вуглекислого газу	0,000038	L/s · W
Потужність освітлення приміщення	10	W/m ²
Потужність електричного обладнання	5	W/m ²
Потужність принтера	200	W
Задана температура на опалення (з 22:00 до 06:00)	15,6	°C
Задана температура на опалення (з 06:00 до 22:00)	21	°C
Задана температура на охолодження (з 22:00 до 06:00)	26,7	°C
Задана температура на охолодження (з 06:00 до 22:00)	24	°C
Алгоритм теплового балансу	Функція теплопередачі через огорожувальні конструкції	
Алгоритм теплового балансу повітря	Метод зворотної різниці третього порядку	
Алгоритм внутрішньої конвекції на поверхні	TARP	
Алгоритм зовнішньої конвекції на поверхні	DOE-2	
Кількість кроків симуляції на годину	4	

на триваліші симуляційні періоди потрібно для вдосконалення підходу, оскільки дає змогу захопити всі складнощі управління HVAC-системою.

Після цього етапу тренування ефективність навченого стратегії керування оцінено шляхом її впровадження та керування моделлю будівлі протягом місяця. Продуктивність контролера глибокого навчання оцінено шляхом аналізу сукупних накопичених винагород. Загалом результати продемонстрували здатність контролера одночасно ефективно оптимізувати кілька цілей, визначених у (3).

Після етапу тренування продуктивність контролера глибокого навчання порівняно з базовою стратегією керування температурою припливного повітря, реалізованою контролером «OutdoorAirReset» Setpoint Manager в EnergyPlus. Цей контролер забезпечує хороший баланс між стратегіями підігріву і охолодження на основі методу зменшення температури зовнішнього повітря. Контролер був налаштований на динамічну корекцію температури припливного повітря на основі зовнішніх умов: з налаштуванням температури подавання повітря 12,8 °C, коли зовнішня температура становить 15,6 °C або нижче, поступово збільшуючись до значення 18,3 °C, коли зовнішня температура досягає 26,7 °C або вище, забезпечуючи ефективне охолодження при збереженні комфорту в різних погодних умовах.

Результати політики керування температурою припливного повітря агентом RL і базової стратегії для першого тижня квітня представлені на рис. 5, де також наведено налаштування термостатів для опалення та охолодження, а також криву температури зовнішнього повітря для порівняння.

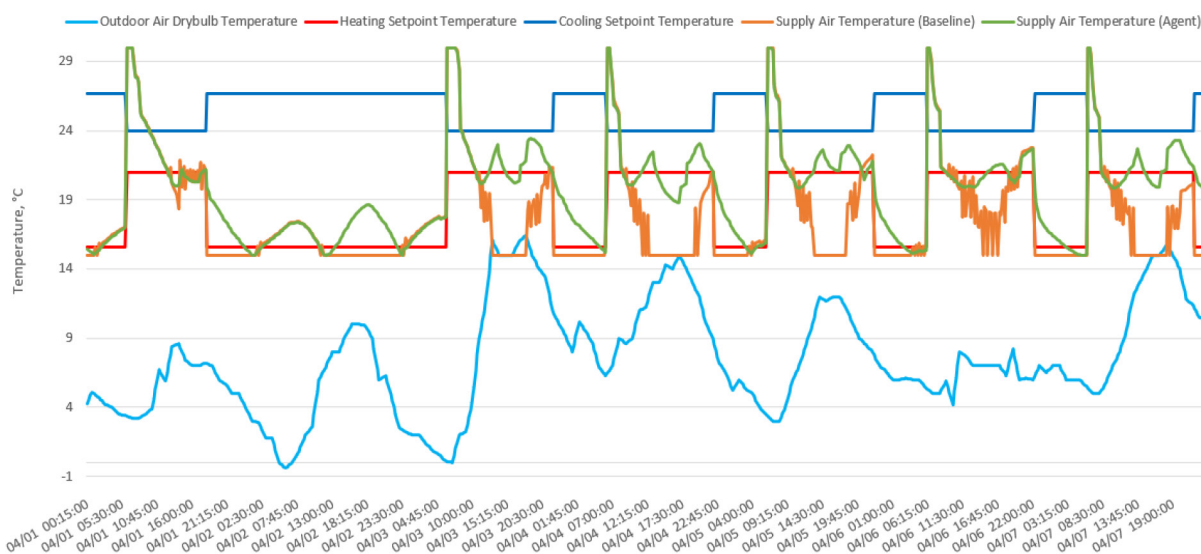


Рис. 5. Політика регулювання температури подавання повітря протягом одного тижня, порівняння між агентом з підкріпленням та базовим контролером

Зауважимо, що 1-го квітня, на початку робочого дня, обидва контролери намагаються підвищити температуру подавання повітря до комфортного для приміщення рівня. Протягом дня температура подавання повітря знижується у зв'язку зі зростанням температури зовнішнього повітря. 3-го квітня протягом робочих годин агент з навчанням з підкріпленням намагається підвищити температуру подавання повітря, щоб підтримувати температуру в приміщенні в межах установлених рівнів нагріву та охолодження. Натомість базовий контролер знижує температуру подавання повітря до мінімального значення, забезпечуючи її в приміщенні на нижній межі комфортного діапазону. Порівняння умов комфорту в приміщенні, включаючи температуру, вологість та концентрацію CO₂ у повітрі, продемонстровано на рис. 6.

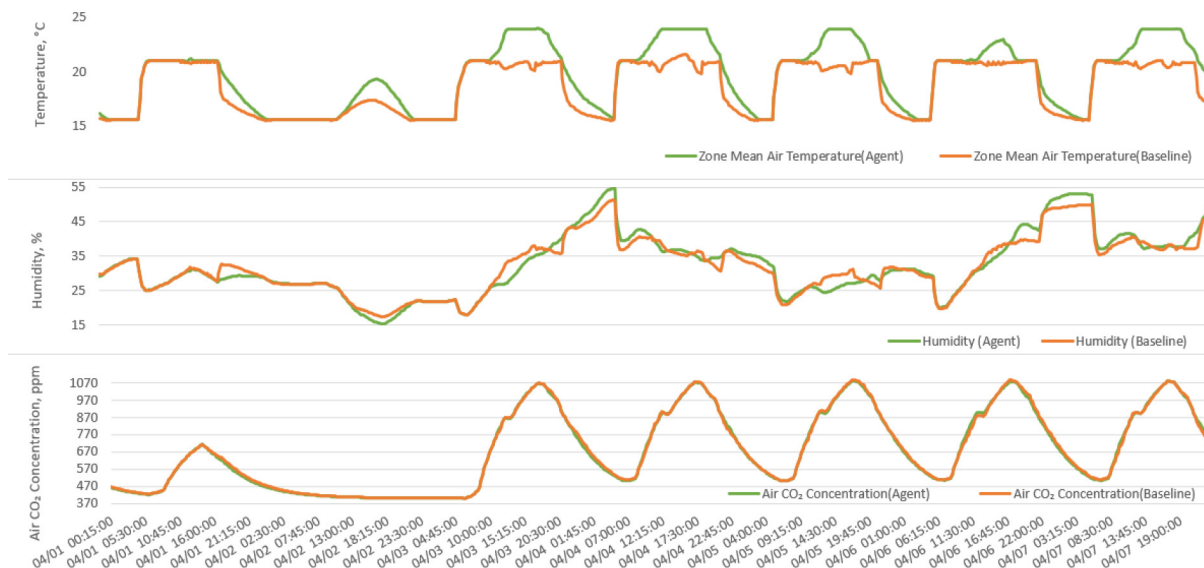


Рис. 6. Порівняння комфорту в приміщенні

Накопичене навантаження на опалення, сформоване агентом з навчанням з підкріпленням та традиційною системою керування на основі методу зменшення температури зовнішнього повітря, продемонстровано на рис. 7. Як видно, алгоритм навчання з підкріпленням демонструє більшу ефективність, забезпечуючи енергоспоживання 962,5 кВт·год енергії за місяць. Це становить скорочення енергоспоживання на 27,8 % порівняно зі стандартним контролером, який забезпечив енергоспоживання 1248,8 кВт·год за той самий період, при цьому зберігаючи належний рівень комфорту в приміщенні. Отримані результати підкреслюють потенціал інтелектуальних систем керування для оптимізації енергоспоживання будівель.

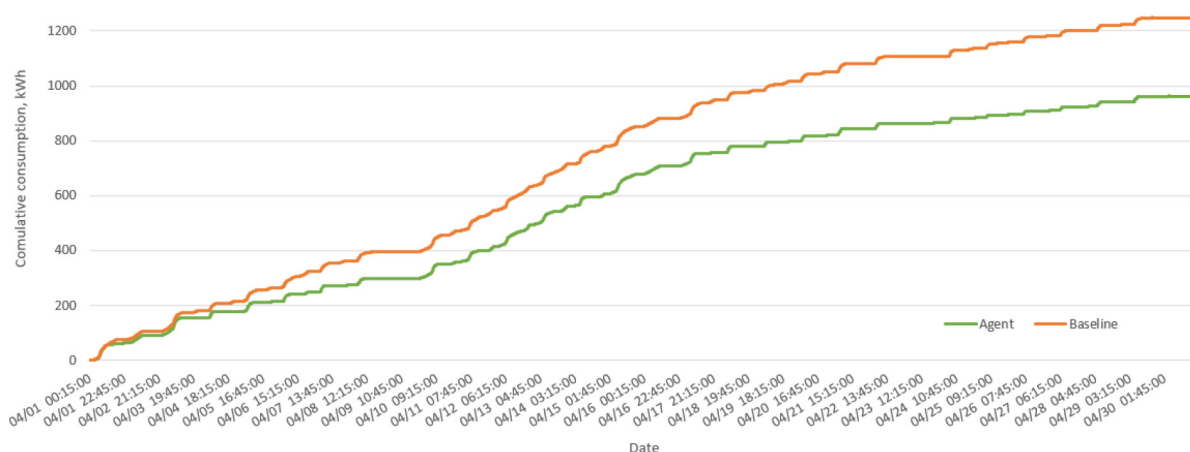


Рис. 7. Накопичене навантаження по опаленню агентом RL та базовим контролером

Висновки

У цьому дослідженні запропоновано інноваційний підхід до управління системою опалення, вентиляції та кондиціювання повітря у будівлях, що передбачає використання інтелектуального контролера з навчанням з підкріпленням, на основі безмодельного алгоритму Proximal Policy Optimization. Запропонований підхід інтегрує

симуляцію енергоспоживання офісної будівлі в EnergyPlus, що забезпечує точний і динамічний опис поведінки системи у різних задачах контролю, зокрема, в регулюванні температури подавання повітря до системи.

Розроблено середовище ко-симуляції Gymnasium для вдосконалення стратегії управління за допомогою алгоритмів навчання з підкріпленням. Ефективність створеного глибокого RL-контролера порівняно з традиційним контролером, який використовує метод коригування за допомогою зовнішньої температури повітря. Результати показали значне покращення енергоефективності, економія енергії склала 27,8 % при збереженні комфортних умов у приміщенні, включаючи температуру, вологість та рівень концентрації CO₂.

Методологія навчання з підкріпленням має потенціал зробити складні стратегії керування доступнішими для впровадження в реальні HVAC-системи. На відміну від традиційного керування на основі моделей, глибоке навчання з підкріпленням не потребує точного математичного представлення фізичної системи. Замість використання складних рівнянь рішення щодо управління приймаються безпосередньо на основі спостережень за зв'язком між виконаними діями та їхніми наслідками для стану системи.

Дослідження підкреслює перспективу поєднання штучного інтелекту та управління системою HVAC, сприяючи впровадженню розумних систем енергоменеджменту в реальних умовах.

Список використаної літератури

1. Walnum H. T., Sartori I., Ward P., Gros S. Demonstration of a low-cost solution for implementing MPC in commercial buildings with legacy equipment. *Applied Energy*. 2025, 380, 125012. <https://doi.org/10.1016/j.apenergy.2024.125012>
2. Perera A. T. D., Kamalaruban P. Applications of reinforcement learning in energy systems. *Renewable and Sustainable Energy Reviews*. 2021, 137, 110618. <https://doi.org/10.1016/j.rser.2020.110618>
3. Cai W., Sawant S., Reinhardt D., Rastegarpour S., Gros S. A Learning-Based Model Predictive Control Strategy for Home Energy Management Systems. *IEEE Access*. 2023, vol. 11, pp. 145264–145280. <https://doi.org/10.1109/ACCESS.2023.3346324>
4. Bayer D., Pruckner M. Enhancing the Performance of Multi-Agent Reinforcement Learning for Controlling HVAC Systems. 2022 IEEE Conference on Technologies for Sustainability (SusTech). 2022, pp. 187–194. <https://doi.org/10.1109/SusTech53338.2022.9794179>
5. Razzano G., Brandi S., Piscitelli M. S., Capozzoli A. Rule extraction from deep reinforcement learning controller and comparative analysis with ASHRAE control sequences for the optimal management of Heating, Ventilation, and Air Conditioning (HVAC) systems in multizone buildings. *Applied Energy*. 2025, vol. 381, 125046. <https://doi.org/10.1016/j.apenergy.2024.125046>
6. Heidari A., Maréchal F., Khovalyg D. An occupant-centric control framework for balancing comfort, energy use and hygiene in hot water systems: A model-free reinforcement learning approach. *Applied Energy*. 2022, vol. 312, 118833. <https://doi.org/10.1016/j.apenergy.2022.118833>
7. Wen Z., O'Neill D., Maei H. Optimal Demand Response Using Device-Based Reinforcement Learning. *IEEE Trans. Smart Grid*. 2015, 6(5), 2312–2324. <https://doi.org/10.1109/TSG.2015.2396993>
8. Huang X., Zhang D., Zhang X. Energy management of intelligent building based on deep reinforced learning. *Alexandria Engineering Journal*. 2021, 60(1), 1509–1517. <https://doi.org/10.1016/j.aej.2020.11.005>
9. Mugumya K. L., Wong J. Y., Chan A., Yip C.-C., Ghazy S. Indoor haze particulate control using knowledge graphs within self-optimizing HVAC control systems. *IOP Conf. Ser.: Earth Environ. Sci.* 2020, 489(1), 012006. <https://doi.org/10.1088/1755-1315/489/1/012006>
10. Kim S., Lim H. Reinforcement Learning Based Energy Management Algorithm for Smart Energy Buildings. *Energies*. 2018, 11(8), 2010. <https://doi.org/10.3390/en11082010>
11. Korkidis P., Dounis A., Kofinas P. Computational Intelligence Technologies for Occupancy Estimation and Comfort Control in Buildings. *Energies*. 2021, 14(16), 4971. <https://doi.org/10.3390/en14164971>
12. Mahapatra C., Moharana A., Leung V. Energy Management in Smart Cities Based on Internet of Things: Peak Demand Reduction and Energy Savings. *Sensors*. 2017, 17(12), 2812. <https://doi.org/10.3390/s17122812>
13. Mocanu E., et al. On-Line Building Energy Optimization Using Deep Reinforcement Learning. *IEEE Trans. Smart Grid*. 2019, 10(4), pp. 3698–3708. <https://doi.org/10.1109/TSG.2018.2834219>
14. Coraci D., et al. A scalable approach for real-world implementation of deep reinforcement learning controllers in buildings based on online transfer learning: The HiLo case study. *Energy and Buildings*. 2025, vol. 329, 115254. <https://doi.org/10.1016/j.enbuild.2024.115254>
15. Spangher L., et al. Prospective Experiment for Reinforcement Learning on Demand Response in a Social Game Framework. *Proceedings of the Eleventh ACM International Conference on Future Energy Systems*. 2020, pp. 438–444. [In English]. <https://doi.org/10.1145/3396851.3402365>
16. Forootani A., Rastegar M., Jooshaki M. An Advanced Satisfaction-Based Home Energy Management System Using Deep Reinforcement Learning. *IEEE Access*. 2022, vol. 10, pp. 47896–47905. <https://doi.org/10.1109/ACCESS.2022.3172327>

17. Park B., Rempel A. R., Lai A. K. L., Chiaramonte J., Mishra S. Reinforcement Learning for Control of Passive Heating and Cooling in Buildings. IFAC-PapersOnLine. 2021, 54(20), pp. 907–912. <https://doi.org/10.1016/j.ifacol.2021.11.287>
18. Almughran O., Slama S. A. B., Zafar B. A. A Reinforcement Learning Approach for Integrating an Intelligent Home Energy Management System with a Vehicle-to-Home Unit. Applied Sciences. 2023, 13(9), 5539. <https://doi.org/10.3390/app13095539>
19. Wang J., Wang Y., Qiu D., Su H., Strbac G., Gao Z. Resilient energy management of a multi-energy building under low-temperature district heating: A deep reinforcement learning approach. Applied Energy. 2025, vol. 378, 124780. <https://doi.org/10.1016/j.apenergy.2024.124780>
20. Tortorelli A., Sabina G., Marchetti B. A Cooperative Multi-Agent Q-Learning Control Framework for Real-Time Energy Management in Energy Communities. Energies. 2024, 17(20), 5199. <https://doi.org/10.3390/en17205199>
21. Vazquez-Canteli J., Kampf J., Nagy Z. Balancing comfort and energy consumption of a heat pump using batch reinforcement learning with fitted Q-iteration. Energy Procedia. 2017, vol. 122, pp. 415–420. <https://doi.org/10.1016/j.egypro.2017.07.429>
22. Cordeiro-Costas M., Villanueva D., Eguía-Oller P., Granada-Alvarez E. Intelligent energy storage management trade-off system applied to Deep Learning predictions. Journal of Energy Storage. 2023, vol. 61, 106784. <https://doi.org/10.1016/j.est.2023.106784>
23. Du Y., Li F., Kurte K., Munk J., Zandi H. Demonstration of Intelligent HVAC Load Management with Deep Reinforcement Learning: Real-World Experience of Machine Learning in Demand Control. IEEE Power and Energy Mag. 2022, 20(3), pp. 42–53. <https://doi.org/10.1109/MPE.2022.3150825>
24. Liu C., Xue Z. Adaptive Optimization Design of Building Energy System for Smart Elderly Care Community Based on Deep Deterministic Policy Gradient. Processes. 2023, 11(7), 2155. <https://doi.org/10.3390/pr11072155>
25. Towers M., et al. Gymnasium: A Standard Interface for Reinforcement Learning Environments. ICLR 2025 Conference, 2025. <https://doi.org/10.48550/ARXIV.2407.17032>
26. Liang E., et al. RLlib: Abstractions for Distributed Reinforcement Learning. Proceedings of the 35th International Conference on Machine Learning. 2018, vol. 80, pp. 3053–3062. <https://doi.org/10.48550/arXiv.1712.09381>

References

1. Walnum, H. T., Sartori, I., Ward, P., & Gros, S. (2025). Demonstration of a low-cost solution for implementing MPC in commercial buildings with legacy equipment. *Applied Energy*, 380, 125012. <https://doi.org/10.1016/j.apenergy.2024.125012>
2. Perera, A. T. D., & Kamalaruban, P. (2021). Applications of reinforcement learning in energy systems. *Renewable and Sustainable Energy Reviews*, 137, 110618. <https://doi.org/10.1016/j.rser.2020.110618>
3. Cai, W., Sawant, S., Reinhardt, D., Rastegarpour, S., & Gros, S. (2023). A learning-based model predictive control strategy for home energy management systems. *IEEE Access*, 11, 145264–145280. <https://doi.org/10.1109/ACCESS.2023.3346324>
4. Bayer, D., & Pruckner, M. (2022). Enhancing the performance of multi-agent reinforcement learning for controlling HVAC systems. In *2022 IEEE Conference on Technologies for Sustainability (SusTech)* (pp. 187–194). IEEE. <https://doi.org/10.1109/SusTech53338.2022.9794179>
5. Razzano, G., Brandi, S., Piscitelli, M. S., & Capozzoli, A. (2025). Rule extraction from deep reinforcement learning controller and comparative analysis with ASHRAE control sequences for the optimal management of Heating, Ventilation, and Air Conditioning (HVAC) systems in multizone buildings. *Applied Energy*, 381, 125046. <https://doi.org/10.1016/j.apenergy.2024.125046>
6. Heidari, A., Maréchal, F., & Khovalyg, D. (2022). An occupant-centric control framework for balancing comfort, energy use and hygiene in hot water systems: A model-free reinforcement learning approach. *Applied Energy*, 312, 118833. <https://doi.org/10.1016/j.apenergy.2022.118833>
7. Wen, Z., O'Neill, D., & Maei, H. (2015). Optimal demand response using device-based reinforcement learning. *IEEE Transactions on Smart Grid*, 6(5), 2312–2324. <https://doi.org/10.1109/TSG.2015.2396993>
8. Huang, X., Zhang, D., & Zhang, X. (2021). Energy management of intelligent building based on deep reinforced learning. *Alexandria Engineering Journal*, 60(1), 1509–1517. <https://doi.org/10.1016/j.aej.2020.11.005>
9. Mugumya, K. L., Wong, J. Y., Chan, A., Yip, C.-C., & Ghazy, S. (2020). Indoor haze particulate control using knowledge graphs within self-optimizing HVAC control systems. *IOP Conference Series: Earth and Environmental Science*, 489(1), 012006. <https://doi.org/10.1088/1755-1315/489/1/012006>
10. Kim, S., & Lim, H. (2018). Reinforcement learning based energy management algorithm for smart energy buildings. *Energies*, 11(8), 2010. <https://doi.org/10.3390/en11082010>
11. Korkidis, P., Dounis, A., & Kofinas, P. (2021). Computational intelligence technologies for occupancy estimation and comfort control in buildings. *Energies*, 14(16), 4971. <https://doi.org/10.3390/en14164971>

12. Mahapatra, C., Moharana, A., & Leung, V. (2017). Energy management in smart cities based on internet of things: Peak demand reduction and energy savings. *Sensors*, 17(12), 2812. <https://doi.org/10.3390/s17122812>
13. Mocanu, E., Mocanu, D. C., Nguyen, P. H., Liotta, A., Webber, M. E., Gibescu, M., & Slootweg, J. G. (2019). On-line building energy optimization using deep reinforcement learning. *IEEE Transactions on Smart Grid*, 10(4), 3698–3708. <https://doi.org/10.1109/TSG.2018.2834219>
14. Coraci, D., Vishwanath, A., Bollinger, L. A., Goy, S., & Hoes, P. J. (2025). A scalable approach for real-world implementation of deep reinforcement learning controllers in buildings based on online transfer learning: The HiLo case study. *Energy and Buildings*, 329, 115254. <https://doi.org/10.1016/j.enbuild.2024.115254>
15. Spangher, L., Gorczyca, M., Luca, M. D., Masood, F. M., Aggarwal, G., Melman, A., Yang, I., & Bayen, A. M. (2020). Prospective experiment for reinforcement learning on demand response in a social game framework. In *Proceedings of the Eleventh ACM International Conference on Future Energy Systems* (pp. 438–444). ACM. <https://doi.org/10.1145/3396851.3402365>
16. Forootani, A., Rastegar, M., & Jooshaki, M. (2022). An advanced satisfaction-based home energy management system using deep reinforcement learning. *IEEE Access*, 10, 47896–47905. <https://doi.org/10.1109/ACCESS.2022.3172327>
17. Park, B., Rempel, A. R., Lai, A. K. L., Chiamonte, J., & Mishra, S. (2021). Reinforcement learning for control of passive heating and cooling in buildings. *IFAC-PapersOnLine*, 54(20), 907–912. <https://doi.org/10.1016/j.ifacol.2021.11.287>
18. Almughran, O., Slama, S. A. B., & Zafar, B. A. (2023). A reinforcement learning approach for integrating an intelligent home energy management system with a vehicle-to-home unit. *Applied Sciences*, 13(9), 5539. <https://doi.org/10.3390/app13095539>
19. Wang, J., Wang, Y., Qiu, D., Su, H., Strbac, G., & Gao, Z. (2025). Resilient energy management of a multi-energy building under low-temperature district heating: A deep reinforcement learning approach. *Applied Energy*, 378, 124780. <https://doi.org/10.1016/j.apenergy.2024.124780>
20. Tortorelli, A., Sabina, G., & Marchetti, B. (2024). A cooperative multi-agent Q-learning control framework for real-time energy management in energy communities. *Energies*, 17(20), 5199. <https://doi.org/10.3390/en17205199>
21. Vazquez-Canteli, J., Kampf, J., & Nagy, Z. (2017). Balancing comfort and energy consumption of a heat pump using batch reinforcement learning with fitted Q-iteration. *Energy Procedia*, 122, 415–420. <https://doi.org/10.1016/j.egypro.2017.07.429>
22. Cordeiro-Costas, M., Villanueva, D., Eguia-Oller, P., & Granada-Alvarez, E. (2023). Intelligent energy storage management trade-off system applied to deep learning predictions. *Journal of Energy Storage*, 61, 106784. <https://doi.org/10.1016/j.est.2023.106784>
23. Du, Y., Li, F., Kurte, K., Munk, J., & Zandi, H. (2022). Demonstration of intelligent HVAC load management with deep reinforcement learning: Real-world experience of machine learning in demand control. *IEEE Power and Energy Magazine*, 20(3), 42–53. <https://doi.org/10.1109/MPE.2022.3150825>
24. Liu, C., & Xue, Z. (2023). Adaptive optimization design of building energy system for smart elderly care community based on deep deterministic policy gradient. *Processes*, 11(7), 2155. <https://doi.org/10.3390/pr11072155>
25. Towers, M., Albrecht, S. V., Bohmer, W., Brockman, G., Cheney, N., Chen, L., Firoiu, V., Greydanus, S., Huang, S., Kappler, D., Klimov, O., Lillicrap, T., Marino, J., Morad, S., Osband, I., Perez, C. F., Racanière, S., & Todorov, E. (2025). Gymnasium: A standard interface for reinforcement learning environments. In *ICLR 2025 Conference*. <https://doi.org/10.48550/ARXIV.2407.17032>
26. Liang, E., Liaw, R., Nishihara, R., Moritz, P., Fox, R., Goldberg, K., Gonzalez, J., Jordan, M., & Stoica, I. (2018). RLlib: Abstractions for distributed reinforcement learning. In *Proceedings of the 35th International Conference on Machine Learning* (Vol. 80, pp. 3053–3062). <https://doi.org/10.48550/arXiv.1712.09381>