

УДК 004.415.3

DOI <https://doi.org/10.32689/maup.it.2025.2.1>

Олександр АРТАМОНОВ

аспірант кафедри інформаційних систем та мереж,
Національний університет «Львівська політехніка»,
oleksandr.o.artamonov@lpnu.ua
ORCID: 0009-0000-8295-733X

Михайло ЛУЧКЕВИЧ

кандидат фізико-математичних наук, доцент,
доцент кафедри інформаційних систем та мереж,
Національний університет «Львівська політехніка»,
mykhailo.m.luchkevych@lpnu.ua
ORCID: 0000-0002-2196-252X

ФОРМАЛІЗАЦІЯ ВИМОГ ДО ІНТЕЛЕКТУАЛЬНИХ СИСТЕМ ПІДТРИМКИ УХВАЛЕННЯ РІШЕНЬ У ЗАДАЧАХ ОПТИМІЗАЦІЇ ХМАРНИХ РЕСУРСІВ

Анотація. У контексті сучасних хмарних обчислювальних середовищ зростає необхідність впровадження автоматизованих механізмів управління ресурсами з метою забезпечення належного рівня продуктивності та оптимізації витрат. Інтелектуальні системи підтримки прийняття рішень виступають ключовим компонентом таких механізмів, оскільки забезпечують аналіз динамічних метрик і прогнозування змін навантаження для адаптивного масштабування.

Мета. Розробка формальної моделі вимог до інтелектуальних систем підтримки прийняття рішень у задачах оптимізації ресурсів хмарної інфраструктури.

Методологія. У межах дослідження здійснено аналіз існуючих підходів до автоскейлінгу та інтелектуальних рішень для керування хмарними середовищами. Застосовано методи системного аналізу для виокремлення категорій вимог та побудови онтологічної моделі, яка включає такі сутності: *CloudResource*, *MonitoringAgent*, *MLModel*, *ScalingPolicy* та *DecisionRecord*. Експериментальна валідація виконана шляхом реалізації прототипу в середовищі *Kubernetes* із застосуванням *Prometheus* для збору метрик.

Наукова новизна. Наукова новизна роботи полягає у комплексній формалізації вимог до інтелектуальних систем підтримки прийняття рішень з урахуванням як функціональних, так і нефункціональних аспектів. Зокрема, до функціональних вимог віднесено моніторинг різномірних метрик, механізми прогнозування, генерацію рекомендацій щодо масштабування та аудит ухвалених рішень.

Висновки. Отримані результати свідчать, що формалізація вимог є необхідною умовою створення надійних і прозорих інтелектуальних систем автоскейлінгу в хмарних середовищах. Запропонована модель забезпечує ефективний баланс між продуктивністю та витратами на інфраструктуру, сприяючи оптимальному використанню ресурсів без погіршення якості обслуговування. У перспективі доцільно розширити онтологію для мультихмарних і безсерверних сценаріїв, розробити методи оперативного оновлення моделей (*Concept Drift*) та вдосконалити механізми *Explainable AI* для забезпечення більш прозорого обґрунтування рекомендацій.

Ключові слова: Інтелектуальні системи підтримки ухвалення рішень, оптимізація хмарних ресурсів, формалізація вимог, автоскейлінг, онтологічна модель.

Oleksandr ARTAMONOV, Mykhailo LUCHKEVYCH. FORMALISATION OF REQUIREMENTS FOR INTELLIGENT DECISION SUPPORT SYSTEMS IN CLOUD RESOURCE OPTIMISATION TASKS

Abstract. In the context of modern cloud computing environments, there is an increasing demand for automated resource management systems to ensure optimal performance and cost efficiency. Intelligent decision support systems are a vital part of these mechanisms, providing analysis of dynamic metrics and forecasting of load changes for adaptive scaling.

Objective. The objective is to develop a formal model of the requirements for intelligent decision support systems for optimising cloud infrastructure resources.

Methodology. This study analyses existing approaches to autoscaling and intelligent solutions for managing cloud environments. System analysis methods were employed to identify categories of requirements and construct an ontological model comprising the following entities: *CloudResource*, *MonitoringAgent*, *MLModel*, *ScalingPolicy* and *DecisionRecord*. Experimental validation was performed by implementing a prototype in a *Kubernetes* environment and using *Prometheus* to collect metrics.

Scientific novelty. This work is scientifically novel due to its comprehensive formalisation of requirements for intelligent decision support systems, considering both functional and non-functional aspects. The functional requirements include monitoring heterogeneous metrics, forecasting mechanisms, generating scaling recommendations and auditing decisions.

Conclusion. The results obtained show that the formalisation of requirements is a prerequisite for the creation of reliable and transparent intelligent autoscaling systems in cloud environments. The proposed model strikes an effective balance

© О. Артамонов, М. Лучкевич, 2025

Стаття поширюється на умовах ліцензії CC BY 4.0

between performance and infrastructure costs, ensuring the optimal use of resources without compromising service quality. Future work should focus on extending the ontology to cover multi-cloud and serverless scenarios, developing methods for rapidly updating models (concept drift) and improving Explainable AI mechanisms to provide more transparent justification for recommendations.

Key words: Intelligent decision support systems, cloud resource optimisation, requirements formalisation, autoscaling, ontological model.

Постановка проблеми. У сучасних інформаційних системах хмарні сервіси дедалі частіше виступають основою для розгортання додатків, зберігання даних та виконання обчислень. Завдяки еластичності та високій доступності, платформи IaaS, PaaS і SaaS [10] забезпечують можливість динамічного масштабування ресурсів відповідно до змін навантаження. Водночас зростання кількості користувачів і ускладнення бізнес-процесів породжує низку завдань щодо оптимального розподілу віртуальних машин, контейнерів або безсерверних функцій з метою мінімізації витрат, збереження належного рівня продуктивності та гарантування необхідних показників якості обслуговування.

Проблема набуває значної ваги як у прикладному, так і в теоретичному аспектах. З одного боку, організації прагнуть зменшити фінансові витрати на утримання хмарної інфраструктури, уникнути надмірного резервування обчислювальних потужностей і скоротити час простою критичних сервісів. З іншого боку, наукова спільнота стикається з рядом невирішених питань: яким чином побудувати модель поведінки хмарної платформи за умов непередбачуваної динаміки навантаження; як інтегрувати алгоритми машинного навчання й штучного інтелекту для адаптивного налаштування конфігурації; яким чином формалізувати вимоги до інтелектуальних систем підтримки ухвалення рішень (ІСПУР) [5], що виконуватимуть функцію «інтелектуального ядра» й забезпечуватимуть аналіз поточного стану системи та генерацію оптимальних рекомендацій щодо резервування й масштабування.

Наявність нечітко окреслених або неповних вимог ускладнює процес розробки ефективних ІСПУР: це призводить до фрагментарного застосування алгоритмів, недостатньої інтеграції з існуючою інфраструктурою й, як наслідок, до зниження загальної ефективності оптимізації хмарних ресурсів. Отже, актуальним завданням є чітке визначення ключових характеристик таких систем, формулювання функціональних та нефункціональних вимог, а також встановлення критеріїв якості, що забезпечать надійну, адаптивну й прозору роботу ІСПУР в умовах динамічних хмарних середовищ.

Аналіз останніх досліджень і публікацій. У останні роки спостерігається значне поглиблення досліджень, присвячених розробці інтелектуальних рішень для управління ресурсами хмарних середовищ. У науковій літературі можна виокремити такі основні напрями досліджень.

Перший напрям стосується методів автоскейлінгу, які базуються на PID-регулюванні та адаптивних алгоритмах зворотного зв'язку [2]. Застосування класичних PID-контролерів дозволяє відносно просто реалізувати механізми автоскейлінгу, однак ефективність таких підходів значною мірою залежить від точного налаштування параметрів контролера [6]. У більшості випадків ця процедура здійснюється вручну або з використанням евристичних методів, що обмежує масштабованість і автоматизацію управління.

Другий напрям пов'язаний із моделями прогнозування завантаження, у яких переважно використовуються методи аналізу часових рядів (зокрема ARIMA [4], Prophet та інші). Ці підходи демонструють обнадійливі результати щодо прогнозування майбутнього навантаження на основі історичних даних, проте їхня точність залежить від стабільності користувацьких патернів та змін зовнішніх умов. Виділення коректних трендів і сезонних складових у даних часто ускладнене неоднорідністю навантаження та наявністю різких, непередбачуваних стрибків у запитах.

Третій напрям охоплює створення інтелектуальних систем підтримки ухвалення рішень із застосуванням методів машинного навчання [1; 7]. У таких системах переважно використовують нейромережеві архітектури (LSTM, GRU, CNN) для прогнозування попиту на ресурси та оптимізації політик розподілу. Окремі дослідники пропонують гібридні рішення, які поєднують евристичні правила з методами глибокого навчання, що дозволяє поєднати точність прогнозування з необхідною швидкістю при ухваленні рішень.

Четвертий напрям присвячений розробці фреймворків для створення інтелектуальних систем підтримки ухвалення рішень у хмарних середовищах [8]. Прикладами таких рішень є OpenStack Telemetry (Ceilometer) та Kubernetes HPA (Horizontal Pod Autoscaler), які надають базові можливості моніторингу та автоскейлінгу. Проте вони не завжди забезпечують достатню гнучкість для реалізації складних сценаріїв ухвалення рішень, тому в сучасних роботах дедалі частіше обговорюють необхідність інтеграції багаторівневого аналізу й методів штучного інтелекту з базовими метриками продуктивності.

П'ятий напрям стосується формалізації вимог до інтелектуальних систем підтримки ухвалення рішень [3]. Деякі публікації пропонують розробку спеціалізованих мов (наприклад, домен-специфічних

мов) для опису політик автоскейлінгу [9], однак питання комплексної формалізації всіх функціональних і нефункціональних вимог ІСПУР досі залишається недостатньо висвітленим.

Аналіз сучасного стану досліджень показує, що більшість робіт зосереджена переважно або на розробці та вдосконаленні алгоритмів прогнозування та адаптивного контролю, або на реалізації ІСПУР у рамках конкретних фреймворків хмарних платформ. Водночас системного узагальнення вимог до таких систем, яке б поєднувало потреби як наукової, так і прикладної спільнот, поки що недостатньо. Саме це обґрунтовує необхідність подальшого поглибленого вивчення та формалізації вимог до інтелектуальних систем підтримки ухвалення рішень у задачах оптимізації хмарних ресурсів.

Метою статті є розробка формальної моделі вимог до інтелектуальних систем підтримки ухвалення рішень у контексті оптимізації хмарних ресурсів. Зокрема, передбачається виконати такі завдання: по-перше, здійснити класифікацію та ієрархізацію функціональних і нефункціональних вимог до ІСПУР для хмарної інфраструктури; по-друге, запропонувати формалізований опис цих вимог із застосуванням онтологічного підходу, що охоплюватиме моделі ресурсів, політик та сценаріїв ухвалення рішень; по-третє, визначити набір базових критеріїв якості (ключових показників ефективності, KPI), які дозволяють оцінювати роботу ІСПУР у режимі реального часу; по-четверте, продемонструвати практичне застосування розробленої моделі вимог шляхом створення прототипу ІСПУР із використанням методів машинного навчання для прогнозування навантаження. Отже, завдання дослідження полягають не лише в теоретичній побудові моделі вимог, а й у перевірці її коректності через прототипну реалізацію та експериментальну валідацію в середовищі Kubernetes з використанням інструментів телеметрії Prometheus.

Виклад основного матеріалу. Інтелектуальна система підтримки ухвалення рішень для оптимізації хмарних ресурсів реалізується у вигляді трирівневої архітектури. Перший рівень – «Шар збору та агрегації метрик» (Monitoring Layer) – забезпечує збір даних за допомогою агентів, які моніторять показники ефективності: завантаження CPU, споживання оперативної пам'яті, пропускну здатність мережі, використання дискового простору, а також метрики рівня додатків (затримка відповіді, довжина черги запитів тощо). Отримана інформація надходить до централізованого сховища часових рядів (Time-Series Database), де зберігається для подальшої обробки.

Другий рівень – «Шар аналітики й прогнозування» (Analytics & Prediction Layer) – відповідає за підготовку даних та побудову прогностичних моделей. Даний шар включає модулі попередньої обробки, які здійснюють нормалізацію, фільтрацію та агрегацію метрик відповідно до заданих часових інтервалів. На очищених даних навчання проходять алгоритми машинного навчання (наприклад, LSTM або Prophet), що передбачають майбутнє навантаження у визначені горизонти прогнозу.

Третій рівень – «Шар ухвалення рішень і застосування політик» (Decision & Control Layer) – генерує рекомендації щодо масштабування та конфігурацій ресурсів на основі прогнозів попереднього рівня. Зокрема, на цьому рівні визначається доцільність горизонтального (збільшення чи зменшення кількості екземплярів) або вертикального (зміни параметрів CPU/memory) масштабування, політики розміщення контейнерів і віртуальних машин, а також адаптивні параметри балансувальників навантаження. Сформовані рішення передаються до компонента оркестрування (наприклад, Kubernetes Controller), який безпосередньо виконує зміни конфігурації інфраструктури.

Вимоги до ІСПУР поділяються на дві основні групи (рис. 1): функціональні та нефункціональні.

Система має забезпечувати збір метрик із різномірних джерел, зокрема з рівня гіпервізора, гостьової операційної системи, контейнерів та додатків, із подальшим агрегуванням даних на різних часових інтервалах (секундному, хвилинному, годинному). Отримані показники зберігаються в історичній базі даних для проведення аналітичної обробки та навчання прогностичних моделей. Інтеграція з фреймворками машинного навчання (наприклад, TensorFlow, PyTorch, scikit-learn) дозволяє здійснювати навчання моделей на історичних даних з урахуванням сезонних та трендових змін, а також адаптивно оновлювати їх у режимі реального часу або потокової обробки нових метрик.

Механізми масштабування реалізуються у двох вимірах: горизонтальному (зміна кількості реплік) та вертикальному (регулювання параметрів CPU і пам'яті). При формуванні рекомендацій щодо масштабування необхідно враховувати існуючі квоти на ресурси – обмеження облікових записів та платіжних планів. Крім того, має бути визначено політики динамічного ребалансування навантаження між обчислювальними ресурсами та забезпечено можливість гібридного режиму ухвалення рішень із участю «людини в циклі» (human-in-the-loop).

Для аудиту та трасування всіх змін інфраструктури передбачено логування відповідних рекомендацій і дій, що застосовуються до цільових об'єктів. Має бути забезпечена можливість відстеження мотивів ухвалення кожного рішення на основі принципів Explainable AI, а також збереження стану системи до й після виконання рекомендацій щодо масштабування.

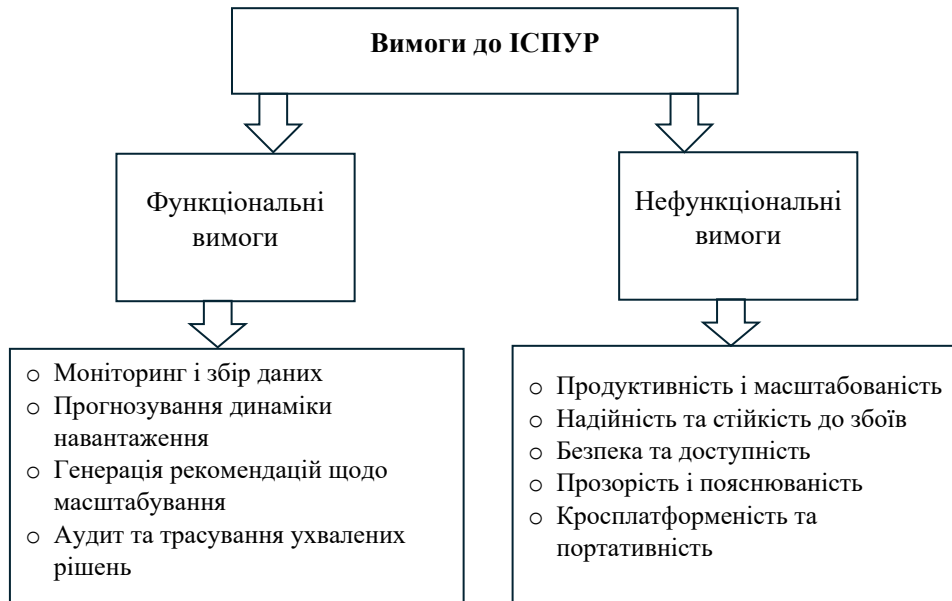


Рис. 1. Вимоги до інтелектуальних систем підтримки ухвалення рішень

Затримка обробки нових метрик і генерації рекомендацій не повинна перевищувати встановленого рівня сервісу (наприклад, одну хвилину з моменту отримання останньої метрики). Підсистеми моніторингу й аналітики повинні бути відмовостійкими та здатними автоматично масштабуватися у відповідь на збільшення навантаження або кількості клієнтських вузлів.

Для забезпечення надійності передбачено збереження консистентності даних у разі втрати зв'язку з окремими вузлами та впровадження механізмів резервування критично важливих компонентів (зокрема кластерного зберігання бази даних і резервних сервісів машинного навчання). Безпеку системи гарантують аутентифікація та авторизація користувачів для доступу до інтерфейсів моніторингу й управління політиками, шифрування каналів зв'язку між агентами моніторингу та сервером збору даних за допомогою TLS, а також ізоляція критичних компонентів з метою запобігання несанкціонованому доступу або виконанню шкідливого коду.

Прозорість ухвалених рішень забезпечується наданням доступу до історії рекомендацій із детальним описом факторів, що впливали на їх формування, а також інтеграцією з інструментами візуалізації (Grafana, Kibana) для аналізу трендів метрик та взаємозв'язків між показниками ресурсів і згенерованими рекомендаціями. Крім того, рішення повинні підтримувати мультихмарні та гібридні сценарії розгортання (AWS, Azure, Google Cloud, on-premises) і забезпечувати уніфікований інтерфейс взаємодії з різними хмарними провайдерами через API-адаптери, що спрощує інтеграцію та гарантує портативність системи.

Для уніфікованого опису вимог до ІСПУР використовується онтологічна модель, яка включає основні сутності: хмарний ресурс, агент моніторингу, модель машинного навчання, політика масштабування та запис ухваленого рішення. Хмарний ресурс характеризується типом (віртуальна машина, контейнер або функція) та відповідними метриками (наприклад, CPUUtilization, MemoryUsage, NetworkIO). Агент моніторингу містить налаштування збору метрик, інтервал опитування та формат передачі даних. Модель машинного навчання описується архітектурою (LSTM, RandomForest, Prophet), набором гіперпараметрів і періодом оновлення. Політика масштабування визначає умови триггеру (наприклад, CPU > 70 % протягом п'яти хвилин), тип масштабування (горизонтальне або вертикальне) та межі кількості ресурсів (мінімальні та максимальні значення). Запис ухваленого рішення зберігає позначку часу (timestamp), набір вхідних метрик, прогнозоване значення навантаження, рекомендовану дію та статус виконання.

Кожна з цих сутностей має власні атрибути та зв'язки. Так, хмарний ресурс містить атрибути resourceID, resourceType і currentMetrics і пов'язаний із агентом моніторингу через зв'язок monitoredBy. Зі свого боку, політика масштабування асоціюється з моделлю машинного навчання за допомогою зв'язку usesModel і визначає перелік цільових ресурсів через зв'язок appliesTo → CloudResource. Така структура моделі дозволяє формалізувати не лише структурні, а й функціональні аспекти доменної логіки: у разі виявлення аномалії навантаження за допомогою відповідної моделі-аномалії (ANOMALY_DETECTOR) система може автоматично активувати резервну політику, залучаючи альтернативні набори ресурсів.

Для об'єктивної оцінки діяльності інтелектуальної системи підтримки ухвалення рішень пропонується застосовувати низку ключових показників ефективності. Першим таким показником є точність прогнозу (Accuracy of Forecast), яка обчислюється на основі середньоквадратичної помилки (RMSE) або середньої абсолютної відносної помилки (MAPE) між передбаченими та фактичними значеннями навантаження. Цей індикатор дозволяє оцінити, наскільки коректно модель машинного навчання формує прогнози й у такий спосіб впливає на якість ухвалених рішень.

Другий показник визначає затримку ухвалення рішення (Decision Latency), що вимірюється як інтервал часу між моментом отримання останнього пакета метрик і часом генерації рекомендації щодо масштабування. Значення цього показника безпосередньо впливає на здатність системи реагувати на зміну навантаження в реальному часі.

Третім індикатором виступає покращення використання ресурсів (Resource Utilization Improvement). Його оцінка здійснюється шляхом порівняння середнього коефіцієнта завантаження CPU та оперативної пам'яті до і після впровадження інтелектуальної системи. Даний показник демонструє ефективність балансування ресурсів та зменшення надмірного резервування.

Четвертим важливим критерієм є зниження витрат (Cost Reduction), яке визначається як процентне співвідношення зменшення витрат на хмарну інфраструктуру у порівнянні з режимом статичного масштабування протягом контрольного періоду. Висока ефективність ухвалених рекомендацій має сприяти зменшенню загальних витрат на обчислювальні ресурси.

Останнім зазначається відповідність рівням обслуговування (Service Level Compliance), що обчислюється як частка випадків перевищення допустимого значення SLA, наприклад, затримки відповіді сервісу понад 200 мс. Моніторинг цього показника дозволяє контролювати якість кінцевого користувацького досвіду й своєчасно виявляти потенційні проблеми.

Постійний супровід зазначених показників дає змогу своєчасно коригувати конфігурацію прогностичних моделей та політик масштабування, а у разі критичних відхилень – оперативно інформувати адміністраторів про необхідність втручання.

Для демонстрації практичної доцільності запропонованої формалізованої моделі вимог розроблено прототип інтелектуальної системи підтримки ухвалення рішень із використанням платформи контейнерної оркестрації Kubernetes, механізму збору та збереження часових рядів метрик Prometheus, засобів візуалізації Grafana і сервера машинного навчання з попередньо навченими моделями LSTM і Prophet для прогнозування навантаження.

Висновки. У статті пропонується формалізація вимог до інтелектуальних систем підтримки ухвалення рішень у задачах оптимізації хмарних ресурсів. Виконано класифікацію функціональних і нефункціональних вимог, визначено ключові критерії якості та розроблено онтологічну модель для формального опису основних сутностей ІСПУР.

Водночас слід зазначити певні обмеження: обрана онтологія потребує адаптації для мультихмарних і гібридних середовищ, оскільки не всі API провайдерів уніфіковано подають метрики. Крім того, експериментальні результати вказують на необхідність розробки механізмів для онлайн-адаптації моделей з урахуванням концептуального зсуву, оскільки змінні патерни навантаження можуть зменшувати точність прогнозування. Також актуальною залишається проблема забезпечення прозорості рекомендацій, що сприятиме довірі з боку адміністраторів, і питання інтеграції безсерверних середовищ (Serverless), де метрики функціонування відрізняються від класичних контейнерних сервісів.

Отже, формалізація вимог до інтелектуальних систем підтримки ухвалення рішень є важливим кроком для створення надійних і прозорих рішень у сфері оптимізації хмарних ресурсів. Подальші зусилля доцільно спрямувати на вдосконалення адаптивного навчання моделей і розширення онтологічної бази з урахуванням специфіки мультихмарних та безсерверних інфраструктур.

Список використаних джерел:

1. Ali R. et al. Intelligent decision support systems – An analysis of machine learning and multicriteria decision-making methods. *Applied Sciences*. 2023. V. 13. №. 22. P. 12426. <https://doi.org/10.3390/app132212426>
2. Burroughs S. et al. Towards autoscaling with guarantees on kubernetes clusters. 2021 IEEE International Conference on Autonomic Computing and Self-Organizing Systems Companion (ACSOS-C). IEEE, 2021. P. 295–296. doi: 10.1109/ACSOS-C52956.2021.00073.
3. Hejase M., Katis A., Mavridou A. Design, formalization, and verification of decision making for intelligent systems. *AIAA SCITECH 2024 Forum*. 2024. P. 2409. <https://doi.org/10.2514/6.2024-2409>
4. Joshi N. S. et al. ARIMA-PID: container auto scaling based on predictive analysis and control theory. *Multimedia Tools and Applications*. 2024. V. 83. №. 9. P. 26369–26386. <https://doi.org/10.1007/s11042-023-16587-0>
5. Onwujekwe G., Weistroffer H. R. Intelligent Decision Support Systems: An Analysis of the Literature and a Framework for Development. *Information Systems Frontiers*. 2025. P. 1–32. <https://doi.org/10.1007/s10796-024-10571-1>

6. Sabuhi M., Mahmoudi N., Khazaei H. Optimizing the performance of containerized cloud software systems using adaptive PID controllers. *ACM Transactions on Autonomous and Adaptive Systems (TAAS)*. 2021. V. 15. №. 3. P. 1–27. <https://doi.org/10.1145/3465630>
7. Saeed A. Machine Learning Models for Intelligent Decision Support Systems. *Journal of AI Range*. 2024. V. 1. №. 1. P. 54–66. URL: <https://www.researchcorridor.org/index.php/jair/article/view/268>
8. Simeone A., Zeng Y., Caggiano A. Intelligent decision-making support system for manufacturing solution recommendation in a cloud framework. *The International Journal of Advanced Manufacturing Technology*. 2021. V. 112. № 3. P. 1035–1050. <https://doi.org/10.1007/s00170-020-06389-1>
9. Tari M. et al. Auto-scaling mechanisms in serverless computing: A comprehensive review. *Computer Science Review*. 2024. V. 53. P. 100650. <https://doi.org/10.1016/j.cosrev.2024.100650>
10. Younis R. et al. A Comprehensive Analysis of Cloud Service Models: IaaS, PaaS, and SaaS in the Context of Emerging Technologies and Trend. 2024 International Conference on Electrical, Communication and Computer Engineering (ICECCE). IEEE, 2024. P. 1–6. doi: 10.1109/ICECCE63537.2024.10823401

Дата надходження статті: 09.06.2025

Дата прийняття статті: 15.06.2025

Опубліковано: 23.09.2025