УДК 004.75:004.8:004.032.26 DOI: 10.31673/2409-7292.2026.028607

Криворучко Віталій Федорович
аспірант кафедри Інженерії програмного забезпечення
Державний університет інформаційно-комунікаційних технологій, Київ, Україна
ORCID: 0009-0004-4247-5722
E-mail: v.kryvoruchko@outlook.com

Садовенко Володимир Сергійович
кандидат фізико-математичних наук, доцент, доцент кафедри Інженерії програмного забезпечення
Державний університет інформаційно-комунікаційних технологій, Київ, Україна
ORCID: 0000-0001-7341-6021
E-mail: v.sadovenko@duikt.edu.ua

ІНТЕЛЕКТУАЛЬНЕ ОЦІНЮВАННЯ ТА ПРЕДИКТИВНА МІНІМЗАЦІЯ ЗАТРИМОК У SERVERLESS-ПЛАТФОРМАХ НА ОСНОВІ ГІБРИДНОЇ НЕЙРОМЕРЕЖЕВОЇ АРХІТЕКТУРИ CNN-LSTM

У статті досліджується проблема холодного старту у Serverless-платформах як одна з основних причин погіршення показників якості обслуговування в умовах нерівномірного та динамічного навантаження. Особливу складність ця проблема становить для середовищ Function-as-a-Service, де стохастичний характер викликів функцій і наявність періодів неактивності призводять до значних затримок ініціалізації обчислювального контексту. Традиційні евристичні політики управління ресурсами, зокрема фіксовані тайм-аути або реактивне масштабування, часто виявляються неефективними за умов різких сплесків навантаження. Запропоновано підхід до предиктивного управління ресурсами Serverless-платформ на основі гібридної нейромережевої архітектури CNN–LSTM, призначеної для прогнозування інтенсивності викликів функцій. Згорткові нейронні мережі використовуються для виділення локальних патернів та короткочасних піків у часових рядах навантаження, тоді як рекурентні LSTM-блоки забезпечують моделювання довгострокових часових залежностей. Отримані прогнози застосовуються для реалізації механізму предиктивного pre-warming, що дозволяє завчасно активувати обчислювальні ресурси та мінімізувати затримки холодного старту. Експериментальне дослідження проведено на реальних трасах виконання Azure Functions із використанням сценаріїв нерівномірного навантаження та періодів неактивності. Результати експериментів показали, що запропонована модель CNN–LSTM забезпечує зниження середньоквадратичної помилки прогнозування на 29,3 % у порівнянні зі стандартною LSTM та дозволяє скоротити кількість холодних стартів у середньому на 32,4 %. Отримані результати підтверджують доцільність використання гібридних нейромережевих підходів для підвищення ефективності та якості обслуговування у сучасних Serverless-системах.

Ключові слова: безсерверні обчислення, Cold Start, CNN-LSTM, прогнозування навантаження, Function-as-a-Service, pre-warming, глибоке навчання.

Вступ

Стрімкий розвиток хмарних обчислень та мікросервісних архітектур зумовив активне впровадження парадигми Serverless-обчислень, зокрема моделі Function-as-a-Service (FaaS), у сучасних розподілених інформаційних системах. Такі платформи, як AWS Lambda, Microsoft Azure Functions та Google Cloud Functions, забезпечують автоматичне масштабування ресурсів та модель оплати за фактичне використання, що робить їх привабливими для високонавантажених веб-застосунків, потокової обробки даних та подійно-орієнтованих систем. Зростання популярності мікросервісної архітектури призвело до того, що розробники все частіше відмовляються від управління сервісами на користь написання бізнес-логіки. Ця зручність приховує технічні виклики, пов’язані з непрогнозованою поведінкою трафіку, де традиційні методи автомасштабування часто реагують з запізненням.

Динамічна природа виділення ресурсів у Serverless-середовищах створює специфічні виклики для забезпечення якості обслуговування (QoS – Quality of Service). Однією з ключових проблем ефективного використання FaaS-платформ є явище холодного старту (Cold Start), яке виникає у випадках, коли функція викликається після періоду не активності і потребує ініціалізації середовища виконання. Це призводить до додаткових затримок, які можуть сягати сотень мілісекунд або навіть секунд, що є критичним для застосунків із

жорсткими вимогами до затримки відповіді [3]. Особливо суттєвим є вплив холодних стартів на довгі хвости розподілу затримок, наприклад, P95-P99 latency, що безпосередньо впливає на дотримання SLA (Service Level Agreement). Окрім безпосередньої затримки виконання коду, холодний старт спричиняє ланцюгову реакцію затримок у залежних сервісах, що може критично вплинути на користувацький досвід у реальному часі. Особливо гостро це питання постає для застосунків Інтернету речей, де мілісекунди затримки можуть призвети до втрати актуальності даних.

З огляду на нерівномірний та часто важко прогнозований характер навантаження у Serverless-системах, традиційні евристичні підходи до управління ресурсами, зокрема статистичні політики Keep-Alive або фіксовані тайм-аути, виявляються малоефективними. Це зумовлює необхідність переходу до адаптивних методів управління, заснованих на аналізі історичних даних та прогнозуванні майбутнього навантаження.

Огляд літератури

Проблема прогнозування навантаження у хмарних обчислювальних середовищах широко розглядається у сучасних наукових дослідженнях. Існуючі підходи можна умовно поділити на кілька основних груп залежно від використовуваного математичного апарату.

Статистичні методи прогнозування, зокрема моделі авторегресії та інтегрованого ковзного середнього (ARIMA), традиційно застосовуються для аналізу часових рядів. У роботі Box і Jenkins [1] показано ефективність ARIMA-моделей для стаціонарних процесів з лінійною динамікою. Проте подальші дослідження [2] у сфері хмарних обчислень вказують на обмежену здатність таких моделей адекватно відображати різкі нелінійні сплески навантаження, характерні для Serverless-середовищ. Хоча класичні методи, такі як експоненційне згладжування, демонструють високу швидкість обчислень, вони часто виявляються безсилими перед різкими, нелінійними змінами навантаження, характерними для FaaS. Їхня основна вада полягає у припущенні стаціонарності процесів, що рідко виконується в умовах реального хмарного трафіку.

Методи класичного машинного навчання, такі як лінійна регресія, вектори підтримки (SVR – Support Vector Regression) та випадкові ліси, застосовуються для прогнозування навантаження у розподілених системах [3]. Їх перевагою є можливість роботи з багатовимірними даними, однак ефективність таких підходів значною мірою залежить від якості ручного відбору ознак, що ускладнює їх використання в умовах динамічного та нерівномірного трафіку.

Суттєвий прогрес у задачах прогнозування часових рядів пов’язаний із використанням глибоких нейронних мереж. Архітектура LSTM (Long Short-Term Memory), запропонована Hochreiter і Schmidhuber [4], продемонструвала високу ефективність у моделюванні довгострокових часових залежностей. Подальші дослідження [5, 6] підтвердили доцільність застосування LSTM-мереж для прогнозування навантаження у хмарних і Edge-системах. На відміну від повнозв’язних мереж, рекурентні архітектури здатні пам’ятати контекст попередніх подій, що є критичним для виявлення добової або тижневої сезонності запитів. Втім, без попередньої обробки вхідних даних такі мережі можуть перенавчатися на шумі, сприймаючи випадкові флуктуації як значущі сигнали.

Разом з тим, аналіз сучасних робіт у сфері оптимізації FaaS-платформ [10] свідчить, що більшість існуючих підходів базуються переважно на рекурентних архітектурах, які, попри здатність до запам’ятовування довгих послідовностей, є вразливими до високочастотного «шуму» в даних моніторингу. У той же час, згорткові нейронні мережі (CNN – Convolutional Neural Network) продемонстрували високу ефективність у задачах просторової фільтрації ознак [7]. Гібридизація цих підходів дозволяє реалізувати механізм, де CNN виступає у ролі інтелектуального препроцесора, що ідентифікує локальні «вибухові» патерни, а LSTM моделює глобальний часовий тренд. Такий синергетичний ефект дозволяє мінімізувати явище концептуального дрейфу [8], коли профіль навантаження змінюється внаслідок оновлення

© Криворучко В. Ф., Садовенко В. С. Інтелектуальне оцінювання та предиктивна мінімзація затримок у serverless-платформах на основі гібридної нейромережевої архітектури CNN-LSTM. Сучасний захист інформації, 2(66), 88–95. https://doi.org/10.31673/2409-7292.2026.028607

сервісів або сезонних коливань. Варто зазначити, що глибокий аналіз архітектури FaaS-платформ, проведений Wang et al. [9], підтверджує: внутрішні політики виділення ресурсів часто залишаються непрозорими (Black Box) для користувача, що робить зовнішнє предиктивне моделювання єдиними ефективним інструментом оптимізації. Дослідження гібридних архітектур у суміжних задачах обробки часових рядів, зокрема робота Karim et al. [10], доводять, що інтеграція згорткових блоків перед рекурентними шарами значно покращує здатність моделі виділяти значущі ознаки з зашумлених даних ще до етапу часового аналізу.

Таким чином, стаття присвячена розробці та дослідженню методу інтелектуального оцінювання продуктивності хмарних платформ із використанням гібридної нейромережевої архітектури CNN-LSTM. Таке поєднання дозволяє реалізувати механізм предиктивного pre-warming (далі – пре-вормінг), що мінімізує час очікування користувача шляхом завчасної підготовки обчислювального контексту на основі інтелектуального прогнозу.

Мета дослідження

Ключовим завданням роботи є розробка та дослідження методів предиктивного управління ресурсами Serverless-платформ на основі гібридної нейромережевої архітектури CNN-LSTM з мінімізацією затримок холодного старту в умовах високої волатильності навантаження.

Актуальність даного дослідження зумовлена стохастичним характером навантаження у хмарних середовищах, що робить статичні стратегії управління ресурсами неефективними. Необхідним є перехід до предиктивних моделей, здатних адаптуватися до динамічної зміни інтенсивності запитів у режимі реального часу.

Об’єктом дослідження є процес функціонування Serverless-платформи в умовах нерівномірного навантаження. Для оцінювання ефективності система розглядається як динамічний об’єкт, стан якого описується інтенсивністю вхідних запитів.

Предметом дослідження є методи та моделі прогнозування навантаження і предиктивного управління ресурсами у FaaS-середовищах.

Математичне моделювання

Математично загальний час виконання функції в Serverless-середовищі можна представити як суму:

$$T_{total} = T_{init} + T_{exec} ,$$

де T_{init} – латентність ініціалізації інфраструктури; T_{exec} – безпосередній час виконання.

Предиктивні методи спрямовані саме на нівелювання складової T_{init} шляхом завчасної активації ресурсів. Проте складність полягає у стохастичній природі викликів: надмірний превормінг призводить до перевитрат бюджету, а недостатній – до порушення параметрів якості обслуговування.

Для моделювання довгострокових часових залежностей у динаміці навантаження використовується рекурентний блок LSTM. Основною перевагою LSTM у порівнянні зі стандартними рекурентними мережами є наявність механізму керуючих вентилів (gates), які дозволяють регулювати потік інформації та запобігати проблемі затухання градієнта під час навчання. Це є критично важливим для задачі прогнозування навантаження, оскільки інтенсивність викликів функцій може залежати як від короткострокових, так і від довгострокових патернів. Вхідний потік запитів до Serverless-функцій представляється як часовий ряд $X = \{x_1, x_2, \dots, x_i\}$, де x_i – кількість викликів функцій за дискретний інтервал часу Δt. Завдання моделі полягає у прогнозуванні значення $\hat{y}_{t+1}$ на наступний крок, або вектор значень на *k* кроків вперед, для прийняття рішення про превентивний запуск.

На першому етапі запропонованої моделі здійснюється екстракція локальних ознак часового ряду за допомогою одновимірних згорткових шарів (Conv1D). Операція згортки для вхідного вектора *x* з використанням фільтра *w* розміром *k* визначається як:

$$h_i^c = \sigma\left(\sum_{j=0}^{k-1} w_j \cdot x_{i+j} + b\right) ,$$

де h_i^c – вихідна ознака; x_{i+j} – вектор вхідних значень у межах навантаження ковзного вікна; w_j – вага фільтра згортки, що оптимізуються під час навчання для розпізнавання специфічних тенденцій; σ – функція активації (ReLU), що відсікає нерелевантні коливання; b – параметр зміщення; k – розмір ядра згортки (кількість аналізованих інтервалів).

Після згортки застосовується шар Max-Pooling для зменшення розмірності та виділення найбільш значущих сигналів:

$$p_i = \max\left(h_{i:i+p}^c\right).$$

Витягнуті згортковим шаром ознаки P відображають локальні патерни навантаження, зокрема короткострокові сплески інтенсивності викликів. Передача цих ознак до LSTM-блоку дозволяє поєднати інформацію про локальні закономірності з довгостроковими часовими залежностями, що забезпечує більш повне представлення динаміки навантаження.

На відміну від стандартних рекурентних мереж, LSTM вирішує проблему затухання градієнта за допомогою механізму воріт. Математична модель комірки LSTM описується системою рівнянь:

1. Вентиль забування (Forget gate) відповідає за визначення тієї частини попереднього стану комірки, яка є нерелевантною для подальшого прогнозування та може бути відкинута. У контексті прогнозування навантаження це дозволяє ігнорувати застарілі або шумові коливання трафіку:

$$f_t = \sigma(W_f \cdot [h_{t-1}, P_t] + b_f);$$

2. Вхідний вентиль (Input gate) визначає, яка нова інформація, отримана з поточного вхідного сигналу, повинна бути збережена у стані комірки. Це дозволяє моделі оновлювати стан комірки з урахуванням новим сплесків та змін у динаміці навантаження:

$$i_t = \sigma(W_i \cdot [h_{t-1}, P_t] + b_i),$$
$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, P_t] + b_C);$$

3. Оновлення стану комірки:

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t;$$

4. Вихідний вентиль (Output gate) формує вихідний прихований стан LSTM, який акумулює найбільш інформативні характеристики часової динаміки навантаження та передається на наступні шари для формування прогнозу:

$$o_t = \sigma(W_o \cdot [h_{t-1}, P_t] + b_o),$$
$$h_t = o_t * \tanh(C_t),$$

де h_t – прихований стан моделі, який містить інформацію про часову динаміку навантаження.

Остаточний прогноз $\hat{y}_{t+1}$ формується повнозв'язним шаром Dense на основні останнього прихованого стану LSTM:

$$\hat{y}_{t+1} = W_y \cdot h_t + b_y.$$

Навчання моделі здійснюється шляхом мінімізації середньоквадратичної помилки (MSE – Mean Squared Error) як цільової функції:

$$L = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2.$$

Використання CNN на першому етапі дозволяє моделі ефективно фільтрувати шум у хмарних логах, тоді як LSTM забезпечує точність прогнозування в умовах складних циклічних навантажень, що властиво сучасним багатокористувацьким платформам.

Для підтвердження ефективності запропонованої гібридної моделі CNN-LSTM було проведено серію експериментів з використанням реальних даних завантаження хмарних платформ.

© Криворучко В. Ф., Садовенко В. С. Інтелектуальне оцінювання та предиктивна мінімзація затримок у serverless-платформах на основі гібридної нейромережевої архітектури CNN-LSTM. Сучасний захист інформації, 2(66), 88–95. https://doi.org/10.31673/2409-7292.2026.028607

В процесі експериментального дослідження було використано підмножину Azure Functions Traces [11], що містить часові ряди викликів окремих Serverless-функцій. Кожен часовий ряд відповідає одній функції та характеризується кількістю викликів за фіксований інтервал часу.

Для забезпечення репрезентативності було відібрано 50 функцій з високою дисперсією навантаження, що характеризуються наявністю як періодів повної неактивності, так і різких сплесків інтенсивності викликів. Саме такі сценарії є найбільш чутливими до виникнення холодного старту.

Дані були агреговані з інтервалом Δt = 1 хвилина, що забезпечує компроміс між часовою роздільною здатністю та стабільністю часових рядів. Перед подачею в нейронну мережу у діапазоні [0,1] було виконану нормалізацію даних:

$$x_{norm} = \frac{x - x_{min}}{x_{max} - x_{min}}.$$

Це дозволило прискорити збіжність алгоритму градієнтного спуску під час навчання. У межах експерименту холодний старт визначався як ситуація, за якої виклик відбувався після періоду неактивності тривалістю, що перевищує порогове значення τ. У такому випадку вважалося, що контейнер виконання було деактивовано, і платформа потребує повторної ініціалізації середовища.

Порогове значення τ обиралося відповідно до стандартної поведінки Azure Functions і становило τ = 15 хвилин. Для кожного інтервалу виклику функції перевірялося, чи відбувався попередній виклик у межах цього інтервалу. На основі прогнозованої інтенсивності викликів модель приймала рішення про превентивне утримання активного контейнера (pre-warming), що дозволяло уникнути повторної ініціалізації у випадку очікуваного зростання навантаження. Для формування навчальних прикладів використовувався підхід ковзного вікна (Sliding Window). Вхідний вектор моделі формувався з останніх L = 30 значень часового ряду, що відповідає 30 хвилинам історичних даних. Цільовим значенням було навантаження на наступному часовому кроці. Таким чином, модель реалізує одно крокове прогнозування, яке використовується для прийняття оперативного рішення щодо активації або деактивації контейнерів виконання.

Гібридна архітектура була реалізована з використанням бібліотеки TensorFlow 2.x із застосуванням високорівневого API Keras. Основні параметри моделі наведено нижче:

- CNN шар: 64 фільтри з розміром ядра k = 3, функція активації ReLU;
- LSTM шар: 100 прихованих нейронів (units);
- Dropout = 0.2 для підвищення узагальнювальної здатності моделі, використання якого обґрунтовано у дослідженні [9];
- Max-Pooling шар для зменшення розмірності та виділення найбільш інформативних ознак;
- повнозв’язний вихідний шар (Dense) для формування остаточного прогнозу.

Навчання моделі здійснювалося з використанням оптимізатора Adam зі швидкістю навчання learning rate = 0.001. Дані було поділено у співвідношенні 80% для навчання та 20% для тестування. Для стабілізації процесу навчання використовувався механізм ранньої зупинки (Early Stopping) на основі валідаційної помилки.

Для кількісного оцінювання точності прогнозування та ефективності платформи використано наступні метрики:

- середньоквадратична помилка (RMSE – Root Mean Square Error), що показує загальну точність передбачення кількості запитів:

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2};$$

- середня абсолютна помилка (MAE – Mean Absolute Error), що відображає середнє арифметичне модулів відхилень:

© Криворучко В. Ф., Садовенко В. С. Інтелектуальне оцінювання та предиктивна мінімзація затримок у serverless-платформах на основі гібридної нейромережевої архітектури CNN-LSTM. Сучасний захист інформації, 2(66), 88–95. https://doi.org/10.31673/2409-7292.2026.028607

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|;$$

- коефіцієнт скорочення холодних стартів (CSRR – Cold Start Reduction Ratio), що відображає безпосередню практичну ефективність методу для кінцевого користувача та хмарного провайдера:

$$CSRR = \frac{C_{baseline} - C_{model}}{C_{baseline}} \cdot 100\%,$$

де $C_{baseline}$ – кількість холодних стартів при стандартній політиці Keep-Alive; C_{model} – кількість холодних стартів при використанні CNN-LSTM пре-вормінгу.

Результати прогнозування запропонованої моделі було порівняно з класичною статистичною моделлю ARIMA, налаштованою індивідуально для кожного часового ряду, та стандартною рекурентною мережею LSTM без згорткових шарів. Вибір моделі ARIMA зумовлений її роллю як еталона класичної статистичної екстраполяції, тоді як порівняння зі стандартною мережею LSTM дозволяє оцінити чистий приріст точності, отриманий завдяки інтеграції згорткових шарів для попередньої фільтрації просторових ознак трафіку.

Таблиця 1

Порівняння точності прогнозів моделей

Модель	RMSE	MAE	CSRR (%)	Час навчання (сек)
ARIMA	0.452	0.381	5.2%	12
Simple LSTM	0.218	0.194	18.7%	145
CNN-LSTM	0.154	0.122	32.4%	168

Згідно Таблиці 1, гібридна модель демонструє найнижчу помилку RMSE = 0.154. Хоча час навчання трохи вищий, ніж у базової LSTM, CNN-блоки дозволяють значно краще ідентифікувати «вибухові» навантаження, які ARIMA повністю ігнорує.

Застосування механізму предуктивного пре-вормінгу на основі CNN-LSTM дозволило знизити середній час очікування ініціалізації T_{wait} на 32.4% порівняно зі стандартними налаштуваннями Azure Functions. Досягнутий показник свідчить про те, що кожен третій потенційний холодний старт було успішно нівельовано на етапі прогнозування. Це безпосередньо корелює з покращенням якості обслуговування (QoS) та підвищенням загальної ефективності хмарної платформи. Застосування гібридної моделі також дозволило знизити середньоквадратичну помилку на 29.3% порівняно зі звичайною LSTM. Головною перевагою CNN-блоків виявилася здатність ідентифікувати періодичні піки навантаження, на яких статистичні моделі ARIMA демонстрували високу похибку.

Висновки

У статті розглянуто проблему холодного старту у Serverless-платформах як одну з ключових причин деградації показників якості обслуговування в умовах нерівномірного та високодинамічного навантаження. Запропоновано підхід до предиктивного управління ресурсами, що базується на використанні гібридної нейромережевої архітектури CNN-LSTM для прогнозування інтенсивності викликів функцій у середовищах Function-as-a-Service.

Проведений аналіз показав, що поєднання згорткових нейронних мереж та рекурентних LSTM-блоків дозволяє одночасно враховувати локальні сплески трафіку та довгострокові часові залежності, характерні для Serverless-навантажень. CNN-компонент виконує функцію інтелектуальної фільтрації шуму та виявлення короткочасних піків, тоді як LSTM забезпечує стабільне моделювання глобальної динаміки процесу.

Експериментальне дослідження, проведене на реальних даних Azure Functions Traces, підтвердило ефективність запропонованого підходу. Гібридна модель CNN-LSTM продемонструвала зниження середньоквадратичної помилки прогнозування на 29.3% порівняно зі стандартною LSTM та суттєву перевагу над класичною статистичною моделлю ARIMA. Використання механізму предиктивного пре-вормінгу на основі отриманих прогнозів

дозволило скоротити кількість холодних стартів у середньому на 32.4%. Наукова новизна роботи полягає у розробці інтегрованого підходу до управління ресурсами Serverless-платформ, який поєднує нейромережеве прогнозування навантаження з формалізованою керуючою логікою холодного старту. На відміну від існуючих підходів, запропонований метод забезпечує адаптацію до динамічних змін профілю навантаження та зменшує негативний вплив концептуального дрейфу.

Практичне значення отриманих знань полягає у можливості застосування розробленої моделі для оптимізації політик pre-warming у хмарних Serverless-платформах, що дозволяє знизити затримки ініціалізації, покращити показники QoS та підвищити ефективність використання обчислювальних ресурсів без істотного збільшення витрат.

Подальші дослідження доцільно спрямувати на інтеграцію багатокрокового прогнозування у механізм прийняття рішень щодо pre-warming, що дозволить враховувати не лише миттєві, а й очікувані майбутні піки навантаження. Крім того, актуальним є дослідження адаптивних стратегій виборі порогового часу неактивності контейнерів на основі статистичних характеристик часових рядів, а також застосування методів пояснюваного машинного навчання для аналізу внеску окремих ознак у прогноз та управлінське рішення.

Перелік посилань

1. Box, G. E. P., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). Time Series Analysis: Forecasting and Control (5th ed.). John Wiley & Sons. URL: https://books.google.com.ua/books?id=De9xBgAAQBAJ.

2. Jonas, E., et al. (2019). Cloud Programming Simplified: A Berkeley View on Serverless Computing. arXiv preprint arXiv:1902.03383. URL: https://arxiv.org/abs/1902.03383.

3. Baldini, I., et al. (2017). Serverless Computing: Current Trends and Open Problems. Research Advances in Cloud Computing, 1–20. URL: https://arxiv.org/abs/1706.03178.

4. Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. Neural Computation, 9(8), 1735–1780. DOI: 10.1162/neco.1997.9.8.1735.

5. Lin, C., & Khazaei, H. (2021). Modeling and Optimization of Performance and Cost of Serverless Applications. IEEE Transactions on Parallel and Distributed Systems, 32(3), 655–669. DOI: 10.1109/TPDS.2020.3028841.

6. Osypanka, P., & Nawrocki, P. (2022). Resource Usage Cost Optimization in Cloud Computing Using Machine Learning. IEEE Transactions on Cloud Computing, 10(3), 2079–2089. DOI: 10.1109/TCC.2020.2987821.

7. O'Shea, K., & Nash, R. (2015). An Introduction to Convolutional Neural Networks. arXiv preprint arXiv:1511.08458. URL: https://arxiv.org/abs/1511.08458.

8. Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep Learning. MIT Press. URL: http://www.deeplearningbook.org.

9. Wang, L., Li, M., Zhang, Y., Ristenpart, T., & Swift, M. (2018). Peeking Behind the Curtains of Serverless Platforms. 2018 USENIX Annual Technical Conference (USENIX ATC 18), 133–146. URL: https://www.usenix.org/conference/atc18/presentation/wang-liang.

10. Karim, F., Majumdar, S., Darabi, H., & Chen, S. (2018). LSTM Fully Convolutional Networks for Time Series Classification. IEEE Access, 6, 1662–1669. DOI: 10.1109/ACCESS.2017.2779939.

11. Shahrad, M., et al. (2020). Serverless in the Wild: Characterizing and Optimizing the Serverless Workload at a Large Cloud Provider. 2020 USENIX Annual Technical Conference (USENIX ATC 20), 205–218. URL: https://www.usenix.org/conference/atc20/presentation/shahrad.

Vitaliy Kryvoruchko
Postgraduate student, Department of Software Engineering
State University of Information and Communication Technologies, Kyiv, Ukraine
ORCID: 0009-0004-4247-5722
E-mail: v.kryvoruchko@outlook.com

Volodymyr Sadovenko
Candidate of Sciences, Associate Professor, Associate Professor of the Department of Software Engineering
State University of Information and Communication Technologies, Kyiv, Ukraine
ORCID: 0000-0001-7341-6021
E-mail: v.sadovenko@duikt.edu.ua

© Криворучко В. Ф., Садовенко В. С. Інтелектуальне оцінювання та предиктивна мінімзація затримок у serverless-платформах на основі гібридної нейромережевої архітектури CNN-LSTM. Сучасний захист інформації, 2(66), 88–95. https://doi.org/10.31673/2409-7292.2026.028607

INTELLIGENT ESTIMATION AND PREDICTIVE MINIMIZATION OF DELAYS IN SERVERLESS PLATFORMS BASED ON HYBRID NEURAL NETWORK ARCHITECTURE CNN-LSTM

The article investigates the problem of cold start in Serverless platforms as one of the main reasons for the deterioration of service quality indicators under conditions of uneven and dynamic load. This problem is particularly difficult for Function-as-a-Service environments, where the stochastic nature of function calls and the presence of periods of inactivity lead to significant delays in the initialization of the computing context. Traditional heuristic resource management policies, in particular fixed timeouts or reactive scaling, often prove ineffective under conditions of sharp load surges. An approach to predictive resource management of Serverless platforms based on a hybrid CNN–LSTM neural network architecture designed to predict the intensity of function calls is proposed. Convolutional neural networks are used to highlight local patterns and short-term peaks in the load time series, while recurrent LSTM blocks provide modeling of long-term time dependencies. The obtained forecasts are used to implement a predictive pre-warming mechanism, which allows to activate computing resources in advance and minimize cold start delays. Experimental research was conducted on real Azure Functions execution routes using scenarios of uneven load and periods of inactivity. The experimental results showed that the proposed CNN–LSTM model provides a 29.3% reduction in the mean square prediction error compared to the standard LSTM and allows to reduce the number of cold starts by an average of 32.4%. The obtained results confirm the feasibility of using hybrid neural network approaches to improve the efficiency and quality of service in modern Serverless systems.

Keywords: serverless computing, Cold Start, CNN-LSTM, load forecasting, Function-as-a-Service, pre-warming, deep learning.

Надійшла до редакції (Received): 20.04.2026
Прийнята до друку (Accepted): 12.06.2026
Опубліковано онлайн (Available online): 25.06.2026

УДК 004.056.5:[004.7:336.717 DOI: 10.31673/2409-7292.2026.029908

Кухарська Наталія Павлівна
кандидат фізико-математичних наук, доцент, доцент кафедри безпеки інформаційних технологій
Національний університет “Львівська політехніка”, Львів, Україна
ORCID: 0000-0002-0896-8361
E-mail: nataliia.p.kukharska@lpnu.ua

Полотай Орест Іванович
кандидат технічних наук, доцент, доцент кафедри управління інформаційною безпекою
Львівський державний університет безпеки життєдіяльності, Львів. Україна
ORCID: 0000-0003-4593-8601
E-mail: orest.polotaj@gmail.com

Павлюк Остап-Йосип Сергійович
здобувач першого (бакалаврського) рівня вищої освіти за спеціальністю 125 “Кібербезпека”
Національний університет “Львівська політехніка”, Львів, Україна
ORCID: 0009-0003-7250-9367
E-mail: ostap-iosyp.pavliuk.kb.2022@lpnu.ua

СИСТЕМНІ АСПЕКТИ ПІДГОТОВКИ ОРГАНІЗАЦІЙ ДО ПРОВЕДЕННЯ АУДИТУ НА ВІДПОВІДНІСТЬ ВИМОГАМ PCI DSS

У статті розглянуто організаційні та практичні аспекти забезпечення готовності організацій до аудиту PCI DSS. Обґрунтовано доцільність сприйняття процесу підтвердження відповідності вимогам стандарту не як формальної щорічної процедури, а як безперервного процесу управління інформаційною безпекою. Проаналізовано типові проблеми, що виникають під час аудиту, зокрема помилки у визначенні сфери застосування, недостатній рівень документування процедур та нерегулярність виконання контрольних заходів. У роботі систематизовано основні етапи аудиту, що проводиться кваліфікованим оцінювачем безпеки (QSA), та визначено ключові напрями підготовки організації до зовнішнього оцінювання або самооцінювання. Окрему увагу приділено питанням оптимізації сфери середовища даних власників карток (CDE), проведенню аналізу

© Криворучко В. Ф., Садовенко В. С. Інтелектуальне оцінювання та предиктивна мінімзація затримок у serverless-платформах на основі гібридної нейромережевої архітектури CNN-LSTM. Сучасний захист інформації, 2(66), 88–95. https://doi.org/10.31673/2409-7292.2026.028607