DOI: https://doi.org/10.15276/ict.03.2026.33
УДК 004.85:621.39

Модель представлення та метод формування ознак потокової інформації мобільної мережі

Ревенко Петро Георгійович[1)]
ORCID: https://orcid.org/0009-0003-7243-7826; petroniy13@stud.op.edu.ua
Стрельцов Олег Васильович[1)]
ORCID: https://orcid.org/0000-0002-4691-5703; streltsov.o.v@op.edu.ua. Scopus Author ID: 57210373473
[1)] Національний університет «Одеська політехніка», пр. Шевченка, 1. Одеса, 65044, Україна

АНОТАЦІЯ

У статті розглянуто задачу формування узгодженого ознакового представлення потокової гетерогенної інформації, що надходить від компонентів мобільної мережі. Актуальність дослідження зумовлена асинхронністю, різною частотою дискретизації, структурною неоднорідністю, пропусками та неоднаковою якістю первинних спостережень, що ускладнює їх безпосереднє використання у моделях машинного навчання. **Метою роботи** є розроблення моделі представлення потокової інформації та методу формування ознакового простору, що забезпечують збереження часової і структурної специфіки окремих джерел та формування інтегрованого представлення стану мобільної мережі. Запропонована модель описує компоненти мережі, їх інформаційні потоки, індивідуальні часові шкали, розмірність первинних спостережень, оцінки якості та міжкомпонентні зв'язки без обов'язкового попереднього приведення всіх потоків до єдиної частоти дискретизації. **Розроблений метод** передбачає синхронізацію даних у спільних часових вікнах, статистичну агрегацію, урахування доступності, якості та свіжості спостережень, формування часових і подієвих характеристик та інтеграцію локальних ознак у вектор стану мережі сталої розмірності. Окремою особливістю методу є збереження інформації про фактичну доступність спостережень, що дозволяє не ототожнювати пропуски з реальними нульовими значеннями мережевих показників. **Експериментальне дослідження** проведено шляхом порівняння семи варіантів ознакового представлення за незмінного алгоритму класифікації. Отримані результати показали підвищення macro-F1 від 0,781 для найпростішого представлення до 0,928 для повного інтегрованого варіанта. Дослідження стійкості до неповноти даних засвідчило меншу деградацію якості запропонованого представлення порівняно з використанням останніх доступних значень та ресемплінгом з інтерполяцією. Це підтверджує **доцільність використання** запропонованого підходу в умовах нерівномірного та неповного надходження мережевої інформації. Результати підтверджують доцільність явного врахування гетерогенності, часової динаміки, якості та доступності потокової інформації під час підготовки даних для задач машинного навчання в мобільних мережах.

Ключові слова: мобільні мережі; машинне навчання; потокові дані; гетерогенні дані; обробка інформації; формування ознак; аналіз часових рядів; телекомунікаційні системи

Для цитування: **Ревенко П. Г., Стрельцов О. В. "Модель представлення та метод формування ознак потокової інформації мобільної мережі". *Інформатика. Культура. Техніка*. 2026; Том 3 № 1 (3): 416–428. DOI: https://doi.org/10.15276/ict.03.2026.33**

Актуальність. Сучасні мобільні мережі є складними розподіленими системами, компоненти яких безперервно формують значні обсяги інформації про мережевий трафік, навантаження, затримки, якість обслуговування, використання ресурсів та події функціонування. Такі дані можуть використовуватися методами машинного навчання для класифікації станів мережевих компонентів, прогнозування навантаження, виявлення аномалій та підтримки керування ресурсами. Водночас мобільне мережеве середовище характеризується високою динамічністю, гетерогенністю джерел та зміною характеристик даних у часі [1].

Інформація з різних компонентів мобільної мережі надходить із неоднаковою частотою, має різну розмірність, часову структуру та рівень повноти. Тому безпосереднє використання окремих поточних значень показників не забезпечує повного відображення стану мережі та може призводити до втрати часових і подієвих залежностей. Для подальшого застосування машинного навчання необхідне узгоджене представлення таких потоків, яке поєднує поточні значення з характеристиками їх зміни, статистичними оцінками, подієвими ознаками та показниками якості даних.

Додаткові вимоги до такого представлення виникають у разі виконання машинної обробки безпосередньо поблизу джерел даних. Edge computing дає змогу переносити частину обчислень із централізованої хмарної інфраструктури до граничних вузлів, зменшуючи затримку та обсяг передавання даних [2]. Поєднання edge-обчислень із моделями штучного

© Ревенко П., Стрельцов О., 2026

Стаття опублікована у відкритому доступі відповідно до умов ліцензії CC BY (https://creativecommons.org/licenses/by/4.0/deed.uk)

інтелекту, своєю чергою, потребує підготовки структурованого представлення даних, придатного для виконання моделей в умовах обмежених обчислювальних ресурсів [3].

Отже, актуальним науково-прикладним завданням є розроблення моделі представлення потокової гетерогенної інформації з компонентів мобільної мережі та методу формування ознак, що забезпечує синхронізацію різнорідних потоків у часових вікнах, урахування статистичних, часових і подієвих характеристик, а також формування інтегрованого представлення стану мережі для подальшого застосування моделей машинного навчання.

Огляд наукових досліджень та публікацій. Значна частина сучасних досліджень, присвячених застосуванню машинного навчання в мобільних мережах, орієнтована на прогнозування мережевого трафіку. У роботі [4] запропоновано підхід до просторово-часового моделювання навантаження стільникової мережі на основі глибокого навчання. Автори показали, що трафік базових станцій характеризується як часовою автокореляцією, так і просторовими залежностями між сусідніми вузлами. Подальший розвиток цього підходу представлено в роботі [5], де для прогнозування міського стільникового трафіку використано графове представлення взаємозв'язків між комірками та враховано просторово-часову структуру спостережень.

У роботі [6] запропоновано модель DeepTP, яка поєднує виділення просторових ознак, представлення зовнішніх факторів та послідовну модель із механізмом уваги для врахування часової динаміки мобільного трафіку. Інший підхід запропоновано в [7], де для прогнозування стільникового трафіку враховуються нелокальні просторові залежності, самоувага та взаємозв'язок просторових і часових ознак різної гранулярності. Результати цих досліджень підтверджують важливість спільного врахування часових і просторових характеристик даних, однак вхідне представлення формується переважно відповідно до конкретної задачі прогнозування трафіку.

Використання прогнозних моделей безпосередньо для керування ресурсами мобільної мережі розглянуто в роботі [8]. Запропонована система DeepCog використовує глибоке навчання для прогнозування потреб у ресурсах окремих мережевих зрізів 5G та подальшого випереджального розподілу пропускної здатності. У роботі [9] досліджено децентралізоване прогнозування мобільного трафіку на edge-вузлах із використанням згорткових і рекурентних нейронних мереж та глибокого перенесення навчання. Такий підхід дає змогу виконувати аналіз ближче до джерел формування інформації та зменшувати потребу в централізованому передаванні даних. Водночас основна увага в цих роботах приділена прогнозній моделі та місцю її виконання, а не побудові універсального представлення різнорідної інформації від компонентів мережі.

Окремий напрям досліджень пов'язаний із виявленням аномалій за показниками функціонування мобільної мережі. У роботі [10] для виявлення аномальних станів стільникової мережі використано глибоку нейронну мережу, навчання якої здійснювалося на великих масивах мережевих даних. У роботі [11] запропоновано просторово-часовий механізм STAD, орієнтований на виявлення аномалій під час моніторингу мобільної мережі. Це підтверджує необхідність урахування не лише поточного значення мережевого показника, але й його часової динаміки та взаємозв'язків з іншими спостереженнями.

У роботі [12] розглянуто виявлення та класифікацію аномалій продуктивності мобільної мережі засобами пояснюваного машинного навчання. Запропонований підхід передбачає підготовку та очищення мережевих даних перед формуванням моделей виявлення аномальних режимів. У дослідженні [13] задача виявлення викидів у ключових показниках мобільної мережі вирішується за допомогою поєднання Temporal Convolutional Network та autoencoder. Особливу увагу приділено сезонності та трендам часових рядів, які можуть істотно впливати на результат виявлення аномалій. Таким чином, ефективність машинного навчання безпосередньо залежить від того, яким чином у вхідному представленні відображено статистичні та часові властивості початкових даних.

Загальні підходи до моделювання багатовимірних часових даних також підтверджують доцільність формування розширеного ознакового представлення. У моделі Temporal Fusion

Transformer [14] передбачено одночасне використання статичних характеристик, часових змінних та різних груп вхідних ознак, а механізми вибору змінних дають змогу визначати їх внесок у результат прогнозування. Модель DeepAR [15] використовує рекурентну архітектуру для формування прогнозів на основі множини пов'язаних часових рядів. Ці підходи створюють ефективні механізми моделювання часових залежностей, проте не визначають спеціалізованої процедури узгодження різночастотних потоків інформації, що надходять від різних компонентів мобільної мережі.

Проведений аналіз показує, що існуючі дослідження окремо вирішують задачі прогнозування мобільного трафіку [4], [5], [6], [7], прогнозування та розподілу ресурсів [8], [9], виявлення аномальних станів [10], [11], [12], [13] і моделювання багатовимірних часових залежностей [14], [15]. При цьому підготовка вхідної інформації здебільшого визначається конкретною прикладною задачею та обраною моделлю машинного навчання. У розглянутих підходах не формується єдине представлення, яке одночасно зберігає приналежність спостережень до компонентів мобільної мережі, їх часові характеристики, статистичну динаміку, подієву інформацію, повноту та якість доступних даних. Це є висновком із порівняння розглянутих постановок задач та способів формування їхніх вхідних даних.

Крім того, за наявності потоків із різними інтервалами надходження даних актуальною залишається задача їх узгодження без обов'язкового приведення всіх джерел до однакової частоти дискретизації. Для подальшого застосування моделей машинного навчання доцільним є формування фіксованого ознакового представлення на основі спільних часових вікон із подальшим виділенням статистичних, часових і подієвих характеристик та явним урахуванням доступності й якості спостережень.

Отже, недостатньо дослідженим залишається питання формування єдиного представлення потокової гетерогенної інформації з компонентів мобільної мережі, придатного для подальшого використання різними моделями машинного навчання. Це зумовлює необхідність розроблення моделі такого представлення та методу формування ознак, які забезпечують узгодження різнорідних потоків у часових вікнах, формування статистичних, часових і подієвих характеристик та побудову інтегрованого представлення поточного стану мобільної мережі.

Метою дослідження є розроблення моделі представлення потокової гетерогенної інформації з компонентів мобільної мережі та методу формування ознакового простору, що забезпечують узгоджену обробку асинхронних різнорідних потоків із урахуванням їх часової структури, якості, доступності, подієвих характеристик і взаємозв'язків між компонентами для подальшого застосування методів машинного навчання.

Модель представлення потокової гетерогенної інформації з компонентів мобільної мережі. Мобільну мережу розглядатимемо як наступну сукупність компонентів:

$$C = \{c_1, c_2, \dots, c_N\}, \tag{1}$$

де c_i – i-й функціональний або обчислювальний компонент мобільної мережі; N – кількість компонентів, інформація з яких використовується у задачі обробки.

Кожному компоненту c_i ставиться у відповідність його тип $g_i \in G$, наприклад RAN, CORE, EDGE або UE/IoT. Тип компонента визначає семантичний контекст доступної інформації та зберігається під час її подальшої обробки.

Компонент c_i генерує потік первинної інформації:

$$X_i = \left\{x_i\left(t_1^{(i)}\right), x_i\left(t_2^{(i)}\right), \dots, x_i\left(t_{K_i}^{(i)}\right)\right\}, \tag{2}$$

а сукупність потоків усіх компонентів визначається як:

$$X = \{X_1, X_2, \dots, X_N\}. \tag{3}$$

Первинне спостереження $x_i(t)$ подається у вигляді чотирьох груп інформації:

$$x_i(t) = [k_i(t), e_i(t), u_i(t), d_i(t)], \tag{4}$$

де $k_{i(t)}$ – мережеві KPI та технічні характеристики; $e_i(t)$ – подієва інформація; $u_i(t)$ – показники використання обчислювальних і системних ресурсів; $d_i(t)$ – показники трафіку та сервісного навантаження.

Для кожного спостереження задається нормована оцінка якості:

$$q_i(t) \in [0,1], \tag{5}$$

а первинний формалізований стан компонента визначається як:

$$z_i(t) = [x_i(t), q_i(t), g_i]. \tag{6}$$

Таким чином, разом із первинними значеннями зберігаються тип джерела та інформація про надійність отриманого спостереження.

На відміну від однорідного табличного представлення, запропонована модель враховує часову та структурну гетерогенність потоків.

Для кожного X_i задається власна множина фактичних часових міток:

$$T_i = \left\{t_1^{(i)}, t_2^{(i)}, \dots, t_{K_i}^{(i)}\right\}. \tag{7}$$

У загальному випадку для різних компонентів:

$$T_i \neq T_j, \;\; i \neq j. \tag{8}$$

Отже, однаковий порядковий номер спостереження у двох потоках не означає, що вони були отримані в один фізичний момент часу. Різною може бути і структура спостережень.

Розмірність первинного вектора визначається як:

$$m_i = \dim(x_i(t)), \tag{9}$$

причому для різних типів компонентів допускається:

$$m_i \neq m_j. \tag{10}$$

Це дає змогу спільно описувати потоки від RAN-, CORE-, EDGE- та інших компонентів без штучного приведення їх первинної структури до однакової розмірності.

Для відображення взаємозв’язків між компонентами вводиться матриця:

$$A = \left[a_{ij}\right]_{N\times N}, \tag{11}$$

де

$$a_{ij} \in [0,1]. \tag{12}$$

Коефіцієнт a_{ij} характеризує релевантність інформації компонента c_j для інтерпретації або обробки стану компонента c_i. Значення a_{ij} можуть визначатися на основі топологічних, сервісних або статистичних зв’язків.

У загальному випадку матриця не є симетричною:

$$a_{ij} \neq a_{ji}. \tag{13}$$

З урахуванням введених складових модель потокової гетерогенної інформації з компонентів мобільної мережі визначається як:

$$D = (C, X, T, m, A), \tag{14}$$

де

$$T = \{T_1, T_2, \dots, T_N\}, \tag{15}$$

$$m = [m_1, m_2, \dots, m_N]. \tag{16}$$

Модель D описує інформаційне середовище мобільної мережі до етапу формування ознак машинного навчання. Вона зберігає належність спостережень до конкретних

компонентів, їх індивідуальні часові шкали, структурну розмірність, тип і якість первинних даних, а також взаємозв'язки між компонентами.

Запропонована модель відрізняється від традиційного підходу, за якого різнорідні потоки попередньо приводяться до уніфікованого представлення. У моделі зберігаються індивідуальні часові шкали, первинна структура потоків, тип і якість спостережень, а також взаємозв'язки між компонентами мобільної мережі. Це дає змогу формально враховувати природну гетерогенність вихідної інформації та використовувати її як основу для подальшої синхронізації потоків і формування ознакового простору для задач машинного навчання.

Метод формування ознак потокової інформації мобільної мережі. Вхідні потоки компонентів мобільної мережі характеризуються різними часовими шкалами, частотою надходження та складом первинних показників, тому їх безпосереднє об'єднання у спільний ознаковий простір є некоректним. У межах дослідження перехід від моделі гетерогенної інформації D до вхідного представлення машинного навчання виконується шляхом часового узгодження потоків, локальної агрегації спостережень і подальшого формування статистичних, часових та подієвих характеристик. При цьому початкова частота дискретизації кожного X_i зберігається і не приводиться примусово до єдиного значення.

Спільний часовий контекст задається послідовністю вікон тривалістю $L > 0$, сформованих із кроком $h > 0$.

Для моменту аналізу τ_n відповідне вікно має вигляд:

$$W_n = (\tau_n - L, \tau_n]. \tag{17}$$

Спостереження компонента c_i, часові мітки яких належать W_n, утворюють локальний блок:

$$B_i(n) = \left\{ z_i\left(t_k^{(i)}\right) | t_k^{(i)} \in T_i \cap W_n \right\}. \tag{18}$$

Отже, однаковий індекс n задає спільний часовий контекст для всіх джерел, тоді як реальні часові мітки $t_k^{(i)}$ усередині блоків залишаються незмінними.

Кількість доступних спостережень визначається як:

$$K_i(n) = |B_i(n)|. \tag{19}$$

Наявність інформації від компонента у поточному вікні фіксується індикатором:

$$v_i(n) = \begin{cases} 1, & K_i(n) > 0, \\ 0, & K_i(n) = 0. \end{cases} \tag{20}$$

Окрім самого факту наявності спостережень, ураховується їхня надійність. Середню оцінку якості даних компонента у вікні визначимо як:

$$\bar{q}_i(n) = \frac{1}{K_i(n)} \sum_{t_k^{(i)} \in T_i \cap W_n} q_i\left(t_k^{(i)}\right), \tag{21}$$

а часову свіжість – через відстань між правою межею вікна та останнім доступним спостереженням:

$$t_i^*(n) = \max(T_i \cap W_n), \tag{22}$$

$$\delta_i(n) = \tau_n - t_i^*(n). \tag{23}$$

Таке подання дозволяє розрізняти три принципово різні ситуації: реальне нульове значення показника, відсутність спостереження та наявність застарілих або недостатньо надійних даних. Доцільність явного врахування масок пропусків та часових інтервалів між спостереженнями також підтверджується результатами досліджень багатовимірних часових рядів з неповними даними [17].

Числові канали $x_{ip}(t)$ характеризуються статистично за всією сукупністю значень у $B_i(n)$. Внесок окремих спостережень коригується нормованими вагами $w_{ik}(n)$, сформованими з урахуванням $q_i(t)$.

Зважене середнє та стандартне відхилення мають вигляд:

$$\mu_{ip}(n) = \sum_{t_k^{(i)} \in T_i \cap W_n} w_{ik}(n) x_{ip}\left(t_k^{(i)}\right), \tag{24}$$

$$\sigma_{ip}(n) = \sqrt{\sum_{t_k^{(i)} \in T_i \cap W_n} w_{ik}(n) \left[x_{ip}\left(t_k^{(i)}\right) - \mu_{ip}(n)\right]^2}. \tag{25}$$

До статистичного блока також включаються мінімальне, максимальне та останнє доступне значення кожного числового каналу. Використання статистичних характеристик часових рядів як ознак є поширеним підходом до побудови інформативного представлення послідовних даних [16]. Окремо формуються характеристики якості, інтенсивності надходження, доступності та свіжості інформації.

У сукупності вони утворюють базове представлення:

$$f_i^0(n) = \left[f_i^S(n), f_i^Q(n)\right], \tag{26}$$

де $f_i^S(n)$ містить статистичні характеристики, а $f_i^Q(n)$ – метаознаки якості та доступності даних.

Опис лише поточного рівня показників не дає змоги відрізнити стабільний стан від швидкої зміни з тим самим поточним значенням. Тому до ознакового простору включається блок часової динаміки.

Для числового каналу оцінюється зміна всередині вікна:

$$\Delta x_{ip}(n) = x_{ip}^+(n) - x_{ip}^-(n), \tag{27}$$

коефіцієнт тренду $b_{ip}(n)$ та зміна середнього рівня відносно попереднього вікна:

$$\Delta\mu_{ip}(n) = \mu_{ip}(n) - \mu_{ip}(n-1). \tag{28}$$

За необхідності цей набір доповнюється першою та другою різницями останніх доступних значень. Часовий блок $f_i^T(n)$ формується виключно з інформації поточного та попередніх вікон, тобто не використовує спостережень, отриманих пізніше за τ_n.

Дискретні події обробляються окремо від числової телеметрії. Для події типу l множина її часових міток у W_n позначається $E_{il}(n)$, а кількість та інтенсивність визначаються співвідношеннями:

$$\mathrm{K}_{\mathrm{il}}^E(n) = |E_{il}(n)|, \qquad \lambda_{il}^E(n) = \frac{K_{il}^E(n)}{L}. \tag{29}$$

Подієвий блок також відображає зміну інтенсивності між сусідніми вікнами, давність останньої події та середній інтервал між повторними подіями. Сукупність цих характеристик позначимо $f_i^E(n)$.

Повне локальне представлення компонента має вигляд:

$$f_i(n) = [f_i^0(n), f_i^T(n), f_i^E(n)]. \tag{30}$$

Вектори компонентів об’єднуються у незмінному порядку, визначеному множиною C:

$$f(n) = [f_1(n), f_2(n), \dots, f_N(n)]. \tag{31}$$

Відсутність даних окремого компонента не змінює структуру $f(n)$, відповідний локальний блок зберігає свою позицію, а стан інформаційної доступності відображається

через введені маски та ознаки якості. Це забезпечує сталу розмірність вхідного простору незалежно від поточного складу доступних спостережень.

На мережевому рівні локальне представлення доповнюється характеристиками взаємозв'язків A, а також показниками повноти та якості спостережуваного стану. Результатом є інтегрований вектор $s(n)$, придатний для використання моделями машинного навчання.

Загальну схему перетворення можна записати як:

$$D \to B(n) \to f_i(n) \to s(n). \tag{32}$$

Особливістю розробленого методу є поєднання часової синхронізації без примусового ресемплінгу, зваженої статистичної агрегації, явного опису якості й доступності інформації та формування часових і подієвих характеристик у межах єдиного ознакового представлення. У результаті асинхронні потоки різної структури перетворюються на вектор сталої розмірності зі збереженням інформації не лише про поточні значення мережевих показників, а й про характер їх зміни та надійність вихідних спостережень.

Експериментальне дослідження. Для експериментальної перевірки запропонованої моделі представлення потокової гетерогенної інформації та методу формування ознак було сформовано синтетичний набір даних, що відтворює функціонування фрагмента мобільної мережі протягом 14 діб. До експериментальної конфігурації включено 24 компоненти: 8 компонентів RAN, 4 компоненти CORE, 4 edge-вузли та 8 агрегованих джерел UE/IoT. Для RAN-компонентів моделювалися показники завантаження радіоресурсів, кількості активних користувачів, пропускної здатності, затримки та втрат пакетів; для CORE – кількість активних сесій, інтенсивність сигнального трафіку, затримка та завантаження ресурсів; для EDGE – завантаження CPU і RAM, довжина черги та час обробки; для UE/IoT – показники якості з'єднання, інтенсивності трафіку та подій handover.

Числові потоки мали різні інтервали спостереження від 2 до 15 сек., тоді як подієва інформація надходила нерегулярно. У дані було введено чотири класи стану мережі: нормальний стан, перевантаження, деградація та критичний стан. Інтервали деградації формувалися спільною зміною декількох взаємопов'язаних показників, що давало змогу перевірити не лише локальні, але й міжкомпонентні ознаки.

Для синхронізації використовувалося часове вікно $L = 60$ сек. із кроком $h = 30$ сек. У результаті було сформовано 40319 часових вікон, з яких 27128 відповідали нормальному стану, 6301 – перевантаженню, 4538 – деградації та 2352 – критичному стану. Відповідно до потокової природи даних вибірку було розділено хронологічно – 24190 вікон використано для навчання, 8063 – для валідації та 8064 – для тестування. Між частинами залишався часовий проміжок, який виключав пряме перекриття первинних спостережень сусідніх вибірок. Така схема відповідає закладеному в методі принципу недопущення витоку інформації між перекривними вікнами.

Як контрольний алгоритм машинного навчання використовувався XGBoost-класифікатор [18]. У всіх експериментах його конфігурація залишалася незмінною (300 дерев, максимальна глибина дерева 6 та швидкість навчання 0,05). Таким чином, зміна результатів визначалася складом вхідного представлення, а не використанням різних алгоритмів класифікації.

Для оцінювання якості використовувалися macro-F1, balanced accuracy та recall критичного класу. Використання цих метрик є доцільним через нерівномірне представлення різних станів мережі в експериментальній вибірці.

На першому етапі проведено дослідження ефективності різних варіантів формування ознакового представлення потокової інформації мобільної мережі (Таблиця 1). Варіант N_1 містив лише останні доступні значення показників у часовому вікні. У N_2 використовувалися статистичні характеристики (середнє, стандартне відхилення, мінімальне, максимальне та останнє значення). Варіант N_3 додатково враховував якість, доступність і свіжість інформації. Представлення N_4 доповнювалося часовими ознаками, а N_5 – подієвими

характеристиками. Варіант N_6 відповідав повному локальному вектору ознак $f_i(n)$, що одночасно включав статистичні, якісні, часові та подієві характеристики. У варіанті N_7 використовувалося повне інтегроване представлення $s(n)$, яке додатково враховувало міжкомпонентні зв'язки та загальномережеві характеристики. Така послідовність відповідає структурі запропонованого методу формування ознак.

Таблиця 1. **Результати порівняння варіантів ознакового представлення**

Варіант	Macro-F1	Balanced accuracy	Recall критичного класу	Час формування, мс
N_1	0,781	0,804	0,712	0,31
N_2	0,812	0,832	0,748	0,74
N_3	0,842	0,858	0,792	1,02
N_4	0,879	0,891	0,836	1,48
N_5	0,861	0,876	0,824	1,31
N_6	0,901	0,910	0,867	1,82
N_7	0,928	0,934	0,901	2,46

Джерело: **розроблено авторами**

Результати показали послідовне підвищення якості класифікації зі збільшенням інформаційної повноти ознакового представлення. Перехід від найпростішого варіанта N_1, що містив лише останні доступні значення показників, до статистичного представлення N_2 підвищив показник macro-F1 з 0,781 до 0,812. Найбільший приріст серед окремих локальних розширень спостерігався для варіанта N_4, у якому базове представлення було доповнено часовими ознаками – покзаник macro-F1 досяг 0,879. Це свідчить про важливість урахування не лише поточного рівня мережевих показників, а й напряму та інтенсивності їх зміни.

Використання подієвих характеристик у варіанті N_5 забезпечило macro-F1 на рівні 0,861. Найкращий результат отримано для інтегрованого представлення N_7, яке додатково враховувало міжкомпонентні зв'язки та загальномережеві характеристики (macro-F1 становив 0,928). Абсолютний приріст відносно N_1 склав 0,147, або 14,7 відсоткового пункту. Одночасно recall критичного класу зріс з 0,712 для N_1 до 0,901 для N_7. Збільшення часу формування ознак з 0,31 мс для N_1 до 2,46 мс для N_7 залишалося незначним порівняно з кроком формування нового вікна $h = 30$ сек. і не створювало суттєвих обмежень для потокової обробки.

Другий етап експерименту було присвячено оцінюванню стійкості ознакового представлення до неповноти вхідної потокової інформації. Для цього у тестовій вибірці контрольовано вилучалася частина первинних спостережень. Досліджувалися рівні пропусків 5, 10, 20, 30 і 40 %. Для кожного значення частки вилучених даних формувалося 10 незалежних випадкових масок, після чого обчислювалися середнє значення Macro-F1 та стандартне відхилення. Такий підхід дозволив зменшити залежність результатів від конкретного розташування пропущених спостережень.

Порівнювалися три способи підготовки вхідної інформації (Таблиця 2). Варіант N_1 використовував лише останні доступні значення показників у часовому вікні. Другий підхід передбачав приведення потоків до спільної часової сітки з подальшою лінійною інтерполяцією пропусків [19]. Варіант N_7 відповідав повному інтегрованому представленню $s(n)$, у якому враховувалися статистичні, часові та подієві характеристики, доступність і якість спостережень, а також міжкомпонентний контекст.

Отримані результати показують нелінійний характер погіршення якості зі збільшенням частки відсутніх даних (Рис. 1). Для варіанта N_1 незначне вилучення 5 % спостережень призвело лише до невеликого зменшення Macro-F1 з 0,781 до 0,772. Проте після перевищення рівня 10-20 % деградація стала суттєвішою – при 30 % відсутніх спостережень значення Macro-F1 знизилося до 0,642, а при 40 % – до 0,579. Це пояснюється тим, що представлення, побудоване на останніх доступних значеннях, поступово втрачає актуальність зі збільшенням кількості пропусків.

Таблиця 2. **Вплив неповноти первинних спостережень на Macro-F1**

Частка вилучених спостережень	N_1	Ресемплінг та інтерполяція	N_7
0 %	$0{,}781 \pm 0{,}004$	$0{,}833 \pm 0{,}004$	$0{,}928 \pm 0{,}003$
5 %	$0{,}772 \pm 0{,}006$	$0{,}828 \pm 0{,}005$	$0{,}926 \pm 0{,}003$
10 %	$0{,}748 \pm 0{,}008$	$0{,}809 \pm 0{,}007$	$0{,}919 \pm 0{,}004$
20 %	$0{,}701 \pm 0{,}011$	$0{,}781 \pm 0{,}009$	$0{,}903 \pm 0{,}005$
30 %	$0{,}642 \pm 0{,}015$	$0{,}724 \pm 0{,}013$	$0{,}881 \pm 0{,}007$
40 %	$0{,}579 \pm 0{,}019$	$0{,}668 \pm 0{,}017$	$0{,}846 \pm 0{,}010$

Джерело: **розроблено авторами**

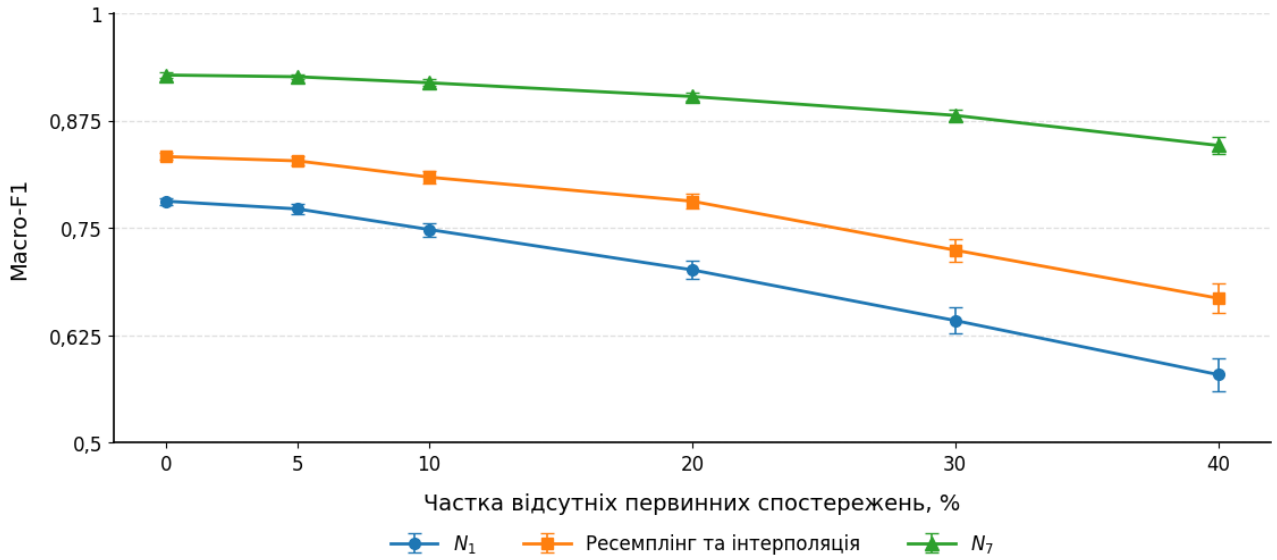

Рис. 1. **Залежність Macro-F1 від частки відсутніх первинних спостережень**

Джерело: **складено авторами**

Ресемплінг з інтерполяцією виявився стійкішим за N_1 при невеликій і помірній частці відсутніх даних. Так, при 20 % вилучених спостережень Macro-F1 становив 0,781 проти 0,701 для N_1. Разом з тим зі збільшенням кількості пропусків ефективність інтерполяції також погіршувалася. За 40 % відсутніх даних показник знизився до 0,668, оскільки значна частина відновлених значень уже визначалася не реальними спостереженнями, а інтерпольованими оцінками.

Найменшу чутливість до неповноти даних продемонстрував варіант N_7. При видаленні 10 % первинних спостережень Macro-F1 зменшився лише з 0,928 до 0,919, а при 20 % – до 0,903. Помітніше погіршення спостерігалося лише при високій частці пропусків. Навіть за відсутності 40 % первинних спостережень Macro-F1 залишався на рівні 0,846.

Відносне погіршення результату визначалося як:

$$D_{F1}(p_{\text{drop}}) = \frac{F1_{\text{macro}}(0) - F1_{\text{macro}}(p_{\text{drop}})}{F1_{\text{macro}}(0)}. \tag{33}$$

При вилученні 30 % спостережень значення D_{F1} становило 17,8 % для N_1, 13,1 % для ресемплінгу з інтерполяцією та 5,1 % для N_7. За 40 % пропусків відносне погіршення складало відповідно 25,9 %, 19,8 % та 8,8 %. Отже, перевага N_7 проявляється не лише у вищому початковому значенні Macro-F1, а й у меншій швидкості деградації результату при зниженні повноти інформації.

Більша стійкість N_7 пояснюється особливостями сформованого ознакового представлення. Відсутність спостереження не замінюється нульовим або штучно

відновленим значенням, а явно відображається через характеристики доступності, якості та свіжості $v_i(n)$, $\bar{q}_i(n)$ та $\eta_i(n)$. Крім того, статистичні характеристики формуються за всіма доступними значеннями в межах часового вікна, а часові та подієві ознаки частково зберігають інформацію про динаміку процесу навіть при втраті окремих первинних спостережень.

Отримані результати підтверджують, що стійкість до неповноти даних залежить не лише від способу заповнення пропусків, а й від структури самого ознакового представлення. Варіант N_1 найбільш чутливий до втрати первинних спостережень, оскільки спирається на обмежений поточний контекст. Ресемплінг з інтерполяцією частково компенсує пропуски, однак при їх значній кількості призводить до збільшення частки штучно сформованих значень. Інтегроване представлення N_7 забезпечує меншу деградацію Macro-F1 завдяки поєднанню статистичних, часових, подієвих і якісних характеристик та збереженню інформації про фактичну доступність первинних потоків.

Висновки. В роботі запропоновано модель представлення потокової гетерогенної інформації з компонентів мобільної мережі та метод формування ознакового простору для подальшого застосування моделей машинного навчання. Підхід враховує різну частоту надходження даних, асинхронність, пропуски, неоднакову якість інформації та структурні відмінності між джерелами. Узгодження потоків виконується у спільних часових вікнах без обов'язкового приведення всіх джерел до єдиної частоти дискретизації.

Розроблений метод забезпечує формування ознакового представлення сталої розмірності, яке поєднує статистичні, часові та подієві характеристики, показники якості, доступності й свіжості інформації, а також міжкомпонентний контекст. Важливо, що відсутність спостереження не ототожнюється з нульовим значенням показника, а явно враховується в ознаковому просторі.

Порівняльне дослідження варіантів ознакового представлення показало, що послідовне розширення представлення підвищує якість класифікації стану мережі. Перехід від останніх доступних значень до статистичних характеристик підтвердив доцільність аналізу всього часового вікна, а врахування якості та свіжості даних дало додатковий приріст. Найбільший внесок серед локальних розширень забезпечили часові ознаки, що свідчить про важливість не лише поточного рівня показників, а й напряму та швидкості їх зміни.

Повне локальне представлення N_6 забезпечило macro-F1 0,901, тоді як інтегроване представлення N_7, що додатково враховувало міжкомпонентні та загальномережеві характеристики, підвищило цей показник до 0,928. Це підтверджує доцільність аналізу стану окремого компонента у контексті процесів, що одночасно відбуваються в інших частинах мобільної мережі.

Дослідження стійкості до неповноти даних також показало перевагу N_7. При вилученні 30 % спостережень відносне погіршення macro-F1 становило 17,8 % для N_1, 13,1 % для ресемплінгу з інтерполяцією та лише 5,1 % для N_7. За 40 % пропусків macro-F1 для N_7 залишався на рівні 0,846. Така стійкість пояснюється тим, що пропуски не приховуються штучно відновленими значеннями, а їх наявність враховується разом із часовою динамікою та іншими характеристиками потоку.

Отримані результати підтверджують, що ефективність запропонованого підходу визначається поєднанням збереження гетерогенності первинних потоків, явного врахування якості та доступності інформації й використання часово-структурного контексту. Сформоване інтегроване представлення $s(n)$ може використовуватися як уніфікований вхід для подальших моделей машинного та глибокого навчання в задачах обробки інформації мобільної мережі.

Конфлікт інтересів: Автори декларують, що не мають конфлікту інтересів, зокрема фінансового, особистого, авторського чи будь-якого іншого характеру, який міг би вплинути на дослідження, а також на результати, опубліковані в цій статті
Фінансування: Дослідження проводилося без фінансової підтримки

Доступність даних: Усі дані доступні в цифровій або графічній формі в основному тексті рукопису

Використання засобів штучного інтелекту (ШІ): Автори підтверджують, що не застосовували інструментальні засоби ШІ для написання цієї роботи

СПИСОК ЛІТЕРАТУРИ

1. Zhang, C., Patras, P. & Haddadi, H. “Deep learning in mobile and wireless networking: A survey”. *IEEE Communications Surveys & Tutorials*. 2019; 21 (3): 2224–2287. DOI: https://doi.org/10.1109/COMST.2019.2904897.

2. Shi, W., Cao, J., Zhang, Q., Li, Y. & Xu, L. “Edge computing: Vision and challenges”. *IEEE Internet of Things Journal*. 2016; 3 (5): 637–646. DOI: https://doi.org/10.1109/JIOT.2016.2579198.

3. Zhou, Z., Chen, X., Li, E., Zeng, L., Luo, K. & Zhang, J. “Edge intelligence: Paving the last mile of artificial intelligence with edge computing”. *Proceedings of the IEEE*. 2019; 107 (8): 1738–1762. DOI: https://doi.org/10.1109/JPROC.2019.2918951.

4. Wang, J., Tang, J., Xu, Z., Wang, Y., Xue, G., Zhang, X. & Yang, D. “Spatiotemporal modeling and prediction in cellular networks: A big data enabled deep learning approach”. *IEEE INFOCOM–IEEE Conference on Computer Communications.* 2017. p. 1–9. DOI: https://doi.org/10.1109/INFOCOM.2017.8057090.

5. Wang, X., Zhou, Z., Xiao, F., Xing, K., Yang, Z., Liu, Y. & Peng, C. “Spatio-temporal analysis and prediction of cellular traffic in metropolis”. *IEEE Transactions on Mobile Computing*. 2019; 18 (9): 2190–2202. DOI: https://doi.org/10.1109/TMC.2018.2870135.

6. Feng, J., Chen, X., Gao, R., Zeng, M. & Li, Y. “DeepTP: An End-to-End neural network for mobile cellular traffic prediction”. *IEEE Network*. 2018; 32 (6): 108–115. DOI: https://doi.org/10.1109/MNET.2018.1800127.

7. Rao, Z., Xu, Y., Pan, S., Guo, J., Yan, Y. & Wang, Z. “Cellular traffic prediction: A deep learning method considering dynamic nonlocal spatial correlation, self-attention, and correlation of spatiotemporal feature fusion”. *IEEE Transactions on Network and Service Management*. 2023; 20 (1): 426–440. DOI: https://doi.org/10.1109/TNSM.2022.3187251.

8. Bega, D., Gramaglia, M., Fiore, M., Banchs, A. & Costa-Pérez, X. “DeepCog: Cognitive network management in sliced 5G networks with deep learning”. *IEEE INFOCOM – IEEE Conference on Computer Communications*. 2019. p. 280–288. DOI: https://doi.org/10.1109/INFOCOM.2019.8737488.

9. Petrella, A., Miozzo, M. & Dini, P. “Mobile traffic prediction at the edge through distributed and deep transfer learning”. *IEEE Access*. 2024; 12: 191288–191303. DOI: 10.1109/ACCESS.2024.3518483.

10. Hussain, B., Du, Q. & Ren, P. “Deep learning-based big data-assisted anomaly detection in cellular networks”. IEEE *Global Communications Conference (GLOBECOM*). 2018. p. 1–6. DOI: https://doi.org/10.1109/GLOCOM.2018.8647366.

11. Dridi, A., Boucetta, C., Hammami, S. E., Afifi, H. & Moungla, H. “STAD: Spatio-temporal anomaly detection mechanism for mobile network management”. *IEEE Transactions on Network and Service Management.* 2021; 18 (1): 894–906. DOI: https://doi.org/10.1109/ TNSM.2020.3048131.

12. Ramírez, J. M., Díez, F., Rojo, P., Mancuso, V. & Fernández-Anta, A. “Explainable machine learning for performance anomaly detection and classification in mobile networks”. *Computer Communications.* 2023; 200: 113–131. DOI: https://doi.org/10.1016/ j.comcom.2023.01.003.

13. Mo, R., Pei, Y., Venkatarayalu, N. V., Nathaniel, P., Premkumar, A. B., Sun, S. & Foo, S. K. K. “Unsupervised TCN-AE-based outlier detection for time series with seasonality and trend for cellular networks”. *IEEE Transactions on Wireless Communications*. 2023; 22 (5): 3114–3127. DOI: https://doi.org/10.1109/TWC.2022.3216004.

14. Lim, B., Arık, S. Ö., Loeff, N. & Pfister, T. “Temporal Fusion Transformers for Interpretable Multi-horizon Time Series Forecasting”. *International Journal of Forecasting*. 2021; 37 (4): 1748–1764. DOI: https://doi.org/10.1016/j.ijforecast.2021.03.012.

15. Salinas, D., Flunkert, V., Gasthaus, J. & Januschowski, T. “DeepAR: Probabilistic forecasting with autoregressive recurrent networks”. *International Journal of Forecasting*. 2020; 36 (3): 1181–1191. DOI: https://doi.org/10.1016/j.ijforecast.2019.07.001.

16. Christ, M., Braun, N., Neuffer, J. & Kempa-Liehr, A. W. “Time series feature extraction on basis of scalable hypothesis tests (tsfresh – A Python package)”. *Neurocomputing*. 2018; 307: 72–77. DOI: https://doi.org/10.1016/j.neucom.2018.03.067.

17. Che, Z., Purushotham, S., Cho, K., Sontag, D. & Liu, Y. “Recurrent neural networks for multivariate time series with missing values”. *Scientific Reports*. 2018; 8: 6085. DOI: https://doi.org/10.1038/s41598-018-24271-9.

18. Moritz, S. & Bartz-Beielstein, T. “imputeTS: Time series missing value imputation in R”. *The R Journal*. 2017; 9 (1): 207–218. DOI: https://doi.org/10.32614/RJ-2017-009.

19. Chen, T. & Guestrin, C. “XGBoost: A scalable tree boosting system”. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016. p. 785–794. DOI: https://doi.org/10.1145/2939672.2939785.

Отримано 24.07.2026
Отримано після перегляду 28.08.2026
Прийнято 03.09.2026

DOI: https://doi.org/10.15276/ict.03.2026.33
UDC 004.85:621.39

A model for representing and a method for feature extraction from streaming mobile network data

Petro H. Revenko[1)]
ORCID: https://orcid.org/0009-0003-7243-7826; petroniy13@stud.op.edu.ua
Oleh V. Streltsov[1)]
ORCID: https://orcid.org/0000-0002-4691-5703; Scopus Author ID: 57210373473; streltsov.o.v@op.edu.ua
[1)] Odesa Polytechnic National University, 1, Shevchenko Av. Odesa, 65044, Ukraine

ABSTRACT

The article addresses the problem of constructing a consistent feature representation of heterogeneous streaming data generated by mobile network components. The relevance of the study is determined by the asynchronous nature of data streams, different sampling rates, structural heterogeneity, missing observations, and varying data quality, all of which complicate their direct use in machine learning models. **The aim** of the study is to develop a model for representing streaming information and a method for feature space construction that preserve the temporal and structural characteristics of individual data sources and provide an integrated representation of the mobile network state. The proposed model describes network components, their data streams, individual time scales, dimensionality of primary observations, data quality estimates, and inter-component relationships without requiring all streams to be preliminarily resampled to a common sampling frequency. The developed **method** includes synchronization of data within common time windows, statistical aggregation, consideration of data availability, quality, and freshness, extraction of temporal and event-based features, and integration of local features into a fixed-dimensional network state vector. A distinctive feature of the method is the explicit preservation of information about observation availability, which prevents missing data from being interpreted as actual zero values of network indicators. The **experimental** study was conducted by comparing seven variants of feature representation using the same classification algorithm. The results showed an increase in macro-F1 from 0.781 for the simplest representation to 0.928 for the full integrated representation. The robustness analysis under incomplete data demonstrated lower performance degradation for the proposed representation compared with the use of the latest available values and resampling with interpolation. The obtained **results** confirm the effectiveness of explicitly accounting for data heterogeneity, temporal dynamics, quality, and availability when preparing streaming mobile network data for machine learning tasks.

Keywords: Mobile networks; machine learning; streaming data; heterogeneous data; information processing; feature extraction; time series analysis; telecommunication systems

ІНФОРМАЦІЯ ПРО АВТОРІВ

Ревенко Петро Георгійович - аспірант, кафедра Комп'ютерних систем. Національний університет «Одеська політехніка», пр. Шевченка, 1. Одеса, 65044, Україна
ORCID: https://orcid.org/0009-0003-7243-7826; petroniy13@stud.op.edu.ua
***Галузь досліджень*:** машинне навчання в мобільних мережах, периферійна інтелектуальна обробка даних, розподілена обробка даних, виявлення аномалій, координація багатоагентних систем, інтелектуальні системи підтримки прийняття рішень

Petro H. Revenko - postgraduate, Department of Computer Systems. Odesa Polytechnic National University, 1, Shevchenko Ave. Odesa, 65044, Ukraine
ORCID: https://orcid.org/0009-0003-7243-7826; petroniy13@stud.op.edu.ua
Research field: machine learning in mobile networks, intelligent edge computing, distributed data processing, anomaly detection, coordination of multi-agent systems, intelligent decision support systems

Стрельцов Олег Васильович – кандидат технічних наук, доцент, доцент кафедри комп`ютерних систем. Національний університет «Одеська політехніка», пр. Шевченка, 1. Одеса, 65044, Україна
***Галузь досліджень*:** комп'ютерні та інформаційно-керуючі системи, цифрова обробка сигналів, синтез і налаштування цифрових фільтрів, сенсорні та інформаційно-вимірювальні системи, бездротові та IoT-мережі, роботизовані та мобільні платформи, інформаційна безпека комп'ютерних систем

Oleh V. Streltsov – Candidate of Engineering Science, Associate Professor, Computer Systems Department. Odesa Polytechnic National University, 1, Shevchenko Ave. Odesa, 65044, Ukraine
ORCID: http://orcid.org/0000-0002-4691-5703; streltsov.o.v@op.edu.ua. Scopus Author ID: 57210373473
Research field: computer and information management systems, digital signal processing, synthesis and tuning of digital filters, sensor and information-measurement systems, wireless and IoT networks, robotic and mobile platforms, information security of computer systems