**DOI: https://doi.org/10.15276/ict.03.2026.30**
**УДК 004.8:004.7**

# Аналіз моделей представлення мережевого трафіку для навчання дифузійних моделей у задачах управління комп'ютерними мережами

**Науменко Роман Іванович**[1)]
ORCID: https://orcid.org/0009-0003-7612-1773; 9530675@stud.op.edu.ua
**Нестеренко Сергій Анатолійович**[1)]
ORCID: https://orcid.org/0000-0002-3757-6594; san@op.edu.ua. Scopus Author Id: 58596482600
[1)] Національний університет «Одеська політехніка», пр. Шевченка, 1. Одеса, 65044, Україна

## АНОТАЦІЯ

У статті досліджено особливості використання моделей представлення мережевого трафіку як навчальних об'єктів дифузійних генеративних моделей у задачах аналізу та управління комп'ютерними мережами. Актуальність роботи зумовлена тим, що реальні мережеві дані часто характеризуються пропусками, нерівномірністю вимірювань, спотвореннями, обмеженнями зберігання та недостатньою представленістю окремих режимів роботи мережі. За таких умов дифузійні моделі можуть використовуватися як для генерації синтетичних зразків, так і для ймовірнісного відновлення відсутніх компонентів, однак спосіб їх застосування залежить від структури сформованого навчального об'єкта. На основі пакетного, потокового, телеметричного, матричного та графового представлень мережевих даних проаналізовано відповідність між типами ознак, часовою і структурною організацією даних та вибором неперервної, дискретної або комбінованої постановки дифузійного моделювання. Формалізовано процес вибору представлення відповідно до властивостей мережевого процесу, необхідних для розв'язання цільової задачі, подальшого формування навчального об'єкта і датасету, постановки дифузійного моделювання, навчання, верифікації та використання отриманого результату. Проаналізовано предметні обмеження відповідних представлень: протокольну та послідовнісну коректність пакетних даних, узгодженість агрегованих характеристик потоків, семантику телеметричних показників і лічильників, просторово-часові залежності матриць трафіку та, залежно від цільової задачі, їх узгодження з маршрутизаційним і ресурсним контекстом, а також відповідність динамічних характеристик топологічній структурі графового представлення. Запропоновано розрізняти гомогенні навчальні об'єкти, сформовані в межах одного представлення, та гетерогенні навчальні об'єкти, що поєднують компоненти декількох представлень мережевих даних. Показано, що гетерогенність навчального об'єкта не є тотожною неоднорідності типів ознак і має розглядатися окремо від використання додаткової відомої інформації як умови генерації або відновлення. Формалізовано постановку ймовірнісного відновлення неповних мережевих спостережень за відомою частиною даних і додатковим контекстом. Обґрунтовано, що відновлений результат слід трактувати як умовно правдоподібний стан, а не як гарантоване відтворення фактично втрачених значень. Для верифікації результатів дифузійної генерації та відновлення узагальнено три взаємодоповнювальні рівні критеріїв: статистичний, предметно-мережевий і прикладний. Отримані результати можуть використовуватися як методологічна основа для вибору представлення мережевих даних, формування навчального об'єкта та постановки дифузійного моделювання відповідно до цільової задачі аналізу або управління комп'ютерною мережею.

**Ключові слова:** мережевий трафік; дифузійні моделі; представлення мережевих даних; генерація мережевого трафіку; відновлення пропущених даних; гомогенні навчальні об'єкти; гетерогенні навчальні об'єкти; управління комп'ютерними мережами

***Для цитування*: Науменко Р. І., Нестеренко С. А. "Аналіз моделей представлення мережевого трафіку для навчання дифузійних моделей у задачах управління комп'ютерними мережами". *Інформатика. Культура. Техніка*. 2026; Том 3, № 1 (3): 367–384. DOI: https://doi.org/10.15276/ict.03.2026.30**

**Актуальність.** Застосування методів машинного навчання в аналізі та управлінні комп'ютерними мережами потребує навчальних даних, які адекватно відображають часові, статистичні, протокольні та структурні властивості мережевого процес.
Водночас придатність таких даних визначається не лише їх обсягом, а й способом представлення, повнотою спостережень та достовірністю зафіксованих характеристик.

У реальних комп'ютерних мережах дані, необхідні для аналізу та управління, нерідко доступні лише частково. Пропуски та нерівномірність вимірювань, втрати або спотворення повідомлень, скидання лічильників, обмеження на зберігання детальних трас і недостатня кількість спостережень рідкісних перевантажень, відмов чи атак призводять до неповноти навчальних наборів і знижують представленість окремих режимів функціонування мережі.

Одним із підходів до генерації та відновлення таких даних є дифузійні генеративні моделі, які навчаються виконувати зворотний до поступового зашумлення процес [1], [2]. Після навчання вони можуть використовуватися як для синтетичної генерації нових зразків, так і, за наявності додаткової відомої інформації (в умовній постановці), для ймовірнісного відновлення відсутніх компонентів за доступними спостереженнями [3]. Остання можливість є особливо актуальною для мережевої телеметрії та інших джерел, у яких пропуски й часткові вимірювання виникають природним чином.

Водночас дифузійна модель відтворює розподіл не абстрактного «мережевого трафіку», а конкретно сформованого навчального об’єкта. У попередній роботі [4] систематизовано пакетне, потокове, телеметричне, матричне та графове представлення, які відрізняються рівнем деталізації, типами ознак і властивостями мережевого процесу, що зберігаються. Тому структура представлення впливає на спосіб дифузійного моделювання, зміст результату генерації або відновлення та вимоги до його предметної коректності.

Навчальний об’єкт може формуватися в межах одного представлення або поєднувати компоненти декількох представлень мережевих даних. У цій роботі такі випадки розглядаються відповідно як гомогенні та гетерогенні навчальні об’єкти. Гетерогенний підхід дає змогу спільно враховувати взаємопов’язані характеристики мережевого стану, зокрема трафікові, топологічні, маршрутизаційні або ресурсні.

Отже, актуальною є задача встановлення відповідності між структурою мережевого навчального об’єкта, способом його дифузійного моделювання, характером генерації або відновлення та критеріями коректності результату. Окремого аналізу потребують способи формування гомогенних і гетерогенних навчальних об’єктів, а також їх використання за умов неповних або частково недостовірних мережевих спостережень.

**Аналіз останніх досліджень і виділення невирішеної частини проблеми.** Теоретичну основу сучасних дифузійних генеративних моделей сформовано в роботах, присвячених дифузійним ймовірнісним моделям з усуненням шуму (Denoising Diffusion Probabilistic Models, DDPM) та генеративному моделюванню на основі градієнта логарифма густини ймовірності, у яких генерація реалізується через навчання зворотного до поступового зашумлення процесу [1], [2]. Одним із практичних обмежень таких моделей є багатоетапність процесу генерації; метод неявних дифузійних моделей з усуненням шуму (Denoising Diffusion Implicit Models, DDIM) дає змогу скоротити кількість кроків генерації [5].

Для мережевих задач важливим методологічним напрямом є дифузійне моделювання багатовимірних часових рядів. Модель TimeGrad орієнтована на ймовірнісне прогнозування, модель CSDI – на умовне відновлення пропущених значень, а модель Diffusion-TS – на генерацію та умовне моделювання часових послідовностей [3], [6], [7]. Хоча ці методи не створювалися спеціально для мережевої телеметрії, вони формують основу для генерації та відновлення часових мережевих показників.

Окремий напрям становить моделювання дискретних і змішаних даних. У дискретних дифузійних моделях D3PM (Structured Denoising Diffusion Models in Discrete State-Spaces) дифузійні переходи задаються безпосередньо для дискретного простору станів [8], тоді як модель TabDDPM орієнтована на об’єкти, що поєднують числові та категоріальні ознаки [9]. Це є суттєвим для пакетних і потокових даних, однак такі загальні підходи самі по собі не забезпечують дотримання специфічних мережевих або протокольних обмежень.

Безпосередні дослідження дифузійного моделювання мережевих даних охоплюють різні рівні їх представлення. Модель NetDiffus використовується для генерації часових характеристик мережевого трафіку [10], модель NetDiff – для формування трас потоків з урахуванням сервісного контексту [11], а модель NetDiffusion – для генерації пакетних трас із контролем протокольної коректності [12]. Для матричного представлення у підході Diffusion-TM дифузійне моделювання застосовується до аналізу та відновлення матриць трафіку, зокрема за неповних спостережень [13]. У моделі STK-Diff просторово-часова генерація мобільного трафіку виконується з використанням додаткового структурованого контексту

[14]. Отже, навіть у межах досліджень, орієнтованих на мережеві дані, постановка дифузійного моделювання суттєво залежить від структури об’єкта, що генерується або відновлюється.

Для графових даних модель DiGress демонструє можливість визначення дифузійного процесу безпосередньо над дискретною графовою структурою та ознаками її компонентів [15]. Ближчою до задач управління комп’ютерними мережами є модель GDPlan, у якій умовна графова дифузія використовується в задачі планування мережі з урахуванням структурних обмежень [16]. Ці роботи підтверджують можливість використання графової структури не лише як форми подання даних, а й як складової умовного генеративного процесу.

Таким чином, розглянуті дослідження підтверджують можливість застосування дифузійних моделей до часових, дискретних, змішаних, матричних, графових і безпосередньо мережевих даних. Водночас кожна з наведених робіт орієнтована переважно на конкретний тип навчального об’єкта або окрему задачу, а вибір способу представлення мережевого трафіку у ній є складовою конкретної постановки.

У попередній роботі [4] систематизовано пакетне, потокове, телеметричне, матричне та графове представлення мережевого трафіку. Водночас у проаналізованих дослідженнях не виявлено узгодженого підходу до розгляду цих представлень саме як навчальних об’єктів дифузійних моделей, який одночасно враховував би типи ознак і відповідний характер дифузійного процесу, використання умовного контексту, зміст генерації або відновлення, предметні мережеві обмеження та критерії оцінювання результату. Окремого аналізу потребують способи формування гомогенних навчальних об’єктів у межах одного представлення та гетерогенних об’єктів, що поєднують компоненти декількох представлень.

**Мета дослідження.** Метою статті є аналіз особливостей використання моделей представлення мережевого трафіку для навчання дифузійних моделей та формалізація процесу вибору представлення, формування навчального об’єкта, постановки дифузійного моделювання, верифікації та використання отриманого результату відповідно до цільової задачі аналізу або управління комп’ютерною мережею. Окрему увагу приділено формуванню гомогенних і гетерогенних навчальних об’єктів та використанню дифузійних моделей за умов неповних або частково недостовірних мережевих спостережень.

Дослідження спирається на пакетне, потокове, телеметричне, матричне та графове представлення мережевого трафіку, детальну формалізацію яких наведено в роботі [4]. У цій статті зазначені представлення використовуються як інформаційна основа для формування навчальних об’єктів дифузійної моделі, а не як повторний об’єкт класифікації. Для застосування запропонованого критерію вибору кожному представленню r ставиться у відповідність множина $P_r$ властивостей мережевого процесу, які воно зберігає.

Для пакетного представлення до $P_r$ належать, зокрема, порядок пакетів, міжпакетні часові інтервали, розміри пакетів, напрямок передачі та протокольні поля. Потокове представлення зберігає агреговані характеристики обміну, зокрема тривалість потоку, кількість пакетів і байтів, протокол та інші характеристики потоку. Для телеметричного представлення суттєвими є часова динаміка показників стану обладнання і трафіку, семантика лічильників та взаємозалежності між вимірюваними характеристиками. Матричне представлення зберігає просторово-часову структуру трафікового попиту між парами джерело–призначення, а графове — топологічні зв’язки та характеристики вузлів і каналів. Наведений перелік узагальнено на основі [4] і використовується в цій роботі безпосередньо для формалізації процедури вибору навчального об’єкта.

Вибір представлення пропонується здійснювати виходячи з властивостей мережевого процесу, які необхідно зберегти для розв’язання цільової задачі.

Нехай Q позначає задачу аналізу або управління мережею, а $P(Q)$ – множину властивостей мережевих даних, необхідних для її розв’язання:

$$P(Q) = \{p_1, p_2, \ldots, p_m\}. \quad (1)$$

де $p_i$ – окрема властивість мережевого процесу, яку необхідно зберегти в навчальних даних для розв'язання задачі $Q$. Таким чином, множина (1) визначає інформаційні вимоги цільової задачі до структури навчального об'єкта.

Для кожного представлення $r$ визначається множина $P_r$ властивостей мережевого процесу, які воно зберігає.

Якщо для певного представлення виконується умова:

$$P(Q) \subseteq P_r, \tag{2}$$

то одного представлення достатньо для формування навчального об'єкта, що відповідає інформаційним вимогам цільової задачі. У цьому випадку формується гомогенний навчальний об'єкт:

$$X^{(hom)} = X_r, \tag{3}$$

де $X_r$ – навчальний об'єкт, сформований відповідно до представлення $r$. Отже, співвідношення (3) визначає випадок, у якому всі необхідні властивості цільової задачі забезпечуються одним представленням.

Якщо умова (2) не виконується для жодного окремого представлення, обирається сукупність представлень $r_1, r_2, \ldots, r_k$, для яких виконується умова:

$$P(Q) \subseteq P_{r1} \cup P_{r2} \cup \ldots \cup P_{rk}. \tag{4}$$

У такому випадку формується гетерогенний навчальний об'єкт:

$$X^{(het)} = \Psi(Ar_1(Xr_1), Ar_2(Xr_2), \ldots, Ar_k(Xr_k)), k \geq 2., \tag{5}$$

де $Ar_i(\cdot)$ позначає операції приведення компонента відповідного представлення до узгодженої форми, а $\Psi(\cdot)$ – спосіб організації узгоджених компонентів у спільному навчальному об'єкті. Операції $Ar_i(\cdot)$ можуть включати приведення часових даних до спільної часової сітки, узгодження частот дискретизації та розмірностей компонентів, нормалізацію числових характеристик, кодування категоріальних ознак, узгодження ідентифікаторів вузлів і каналів, а також формування масок доступності даних. Оператор $\Psi(\cdot)$ у співвідношенні (5) не обов'язково означає безпосередню конкатенацію компонентів: залежно від структури використаних представлень він може задавати їх спільну часову, матричну або графово-структуровану організацію.

При цьому серед $r_1, r_2, \ldots, r_k$ мають бути представлені щонайменше два різні типи представлення. Наприклад, у межах одного навчального об'єкта можуть поєднуватися телеметричні показники з графовими характеристиками вузлів і каналів або матриця трафіку з маршрутизаційними чи ресурсними характеристиками мережі. Співвідношення (2)-(5) не задають числової оптимізаційної процедури, а формалізують принцип вибору та узгодження структури мережевих даних відповідно до інформаційних вимог цільової задачі.

У межах цієї роботи терміни «гомогенний» та «гетерогенний» характеризують склад представлень, з яких сформовано навчальний об'єкт, а не однорідність типів його ознак. Тому, наприклад, пакетний навчальний об'єкт може одночасно містити числові, бінарні та категоріальні компоненти, але за використання лише пакетного представлення залишається гомогенним. Типи ознак визначають особливості дифузійного процесу, тоді як гомогенність або гетерогенність характеризує склад використаних представлень.

Під умовним дифузійним моделюванням у цій роботі розуміється постановка, за якої генерація або відновлення виконується з урахуванням додаткової відомої інформації c, що задає умови формування результату. До такого контексту можуть належати параметри режиму роботи мережі, часові характеристики, умови цільового сценарію, топологія, маршрутизація або доступна частина спостереження.

Для однозначного розмежування компонента навчального об'єкта та умовного контексту пропонується використовувати характер його участі в дифузійному процесі. Якщо компонент

входить до $Xr$ або $X^{(het)}$, піддається відповідному прямому дифузійному перетворенню та має бути згенерований або відновлений у зворотному процесі, він розглядається як складова навчального об'єкта. Якщо відповідна інформація залишається відомою під час генерації або відновлення та використовується лише для обумовлення зворотного процесу, вона розглядається як умовний контекст c.

Тому одна й та сама за своєю природою інформація може виконувати різну роль залежно від постановки задачі. Наприклад, фіксована топологія мережі може використовуватися як умовний контекст під час генерації або відновлення динамічних характеристик вузлів і каналів. Якщо ж сама топологічна структура входить до об'єкта, який модель повинна генерувати або відновлювати, вона стає компонентом навчального об'єкта. Отже, використання умовного контексту саме по собі не є підставою для віднесення навчального об'єкта до гетерогенного.

Після вибору представлення або їх сукупності визначається структура навчального об'єкта $X_R$, де $R$ позначає вибране представлення або набір представлень. На основі сформованих навчальних об'єктів створюється навчальний датасет $D_R$. Його попередня обробка визначається структурою даних і може включати нормалізацію числових характеристик, кодування категоріальних компонентів, обробку пропусків, узгодження часових масштабів та, за необхідності, формування масок спостережень.

Типи компонентів $X_R$ визначають вибір неперервного, дискретного або комбінованого дифузійного процесу. Потреба у використанні умовного контексту та його склад, своєю чергою, визначаються цільовою задачею генерації або відновлення. Сукупність цих рішень утворює постановку дифузійного моделювання $F_R$, відповідно до якої виконується навчання моделі $M_\theta$.

Узагальнений процес побудови, навчання та використання дифузійної моделі можна подати як

$$Q \rightarrow P(Q) \rightarrow R \rightarrow X_R \rightarrow D_R \rightarrow F_R \rightarrow M_\theta \rightarrow Y \rightarrow V \rightarrow U., \quad (6)$$

де $Q$ – цільова задача аналізу або управління мережею; $P(Q)$ – множина властивостей мережевого процесу, необхідних для її розв'язання; $R$ – вибране представлення або набір представлень; $X_\mathrm{R}$ – структура сформованого навчального об'єкта; $D_\mathrm{R}$ – навчальний датасет; $F_R$ – постановка дифузійного моделювання, що охоплює тип дифузійного процесу та використання умовного контексту; $M_\theta$ – навчена дифузійна модель з параметрами $\theta$; $Y$ – результат генерації або ймовірнісного відновлення; $V$ – процедура верифікації результату; $U$ – використання отриманих даних у цільовій задачі аналізу або управління мережею.

Послідовність (6) узагальнює запропонований методологічний процес від постановки цільової задачі до використання результату генерації або відновлення. Її графічне представлення наведено на Рис. 1. Вихідною точкою процесу є цільова задача $Q$, яка визначає множину необхідних властивостей $P(Q)$. Відповідно до умов (2) або (4) обирається одне представлення або їх сукупність, після чого відповідно до (3) або (5) формується гомогенний чи гетерогенний навчальний об'єкт. На його основі створюється датасет, визначається постановка дифузійного моделювання та виконується навчання моделі. Результат генерації або відновлення перед використанням у цільовій задачі проходить верифікацію.

Для демонстрації застосування формалізованого процесу розглянемо задачу ймовірнісного відновлення пропущених значень мережевої телеметрії. У цьому випадку цільова задача $Q$ полягає у відновленні відсутніх показників за доступною частиною багатовимірного часового ряду. До множини $P(Q)$ належать збереження часової динаміки показників, міжознакових залежностей і семантики телеметричних лічильників. Зазначені властивості забезпечуються телеметричним представленням, тому виконується умова (2), а відповідно до (3) формується гомогенний навчальний об'єкт $X_R$ у вигляді багатовимірних часових послідовностей. Навчальний датасет $D_R$ формується з часових вікон телеметрії та масок доступності спостережень. Оскільки основні компоненти такого об'єкта є числовими,

для $F_R$ доцільною є неперервна умовна постановка дифузійного моделювання, у якій доступні значення, маска спостережень і, за потреби, часові характеристики використовуються як умовний контекст. Результатом $Y$ є один або декілька ймовірнісно відновлених варіантів відсутніх компонентів. На етапі $V$ перевіряються похибка на контрольних прихованих значеннях, збереження часових і міжознакових залежностей та предметна коректність телеметричних показників, після чого результат може використовуватися $U$ для подальшого аналізу стану або підтримки задач управління мережею.

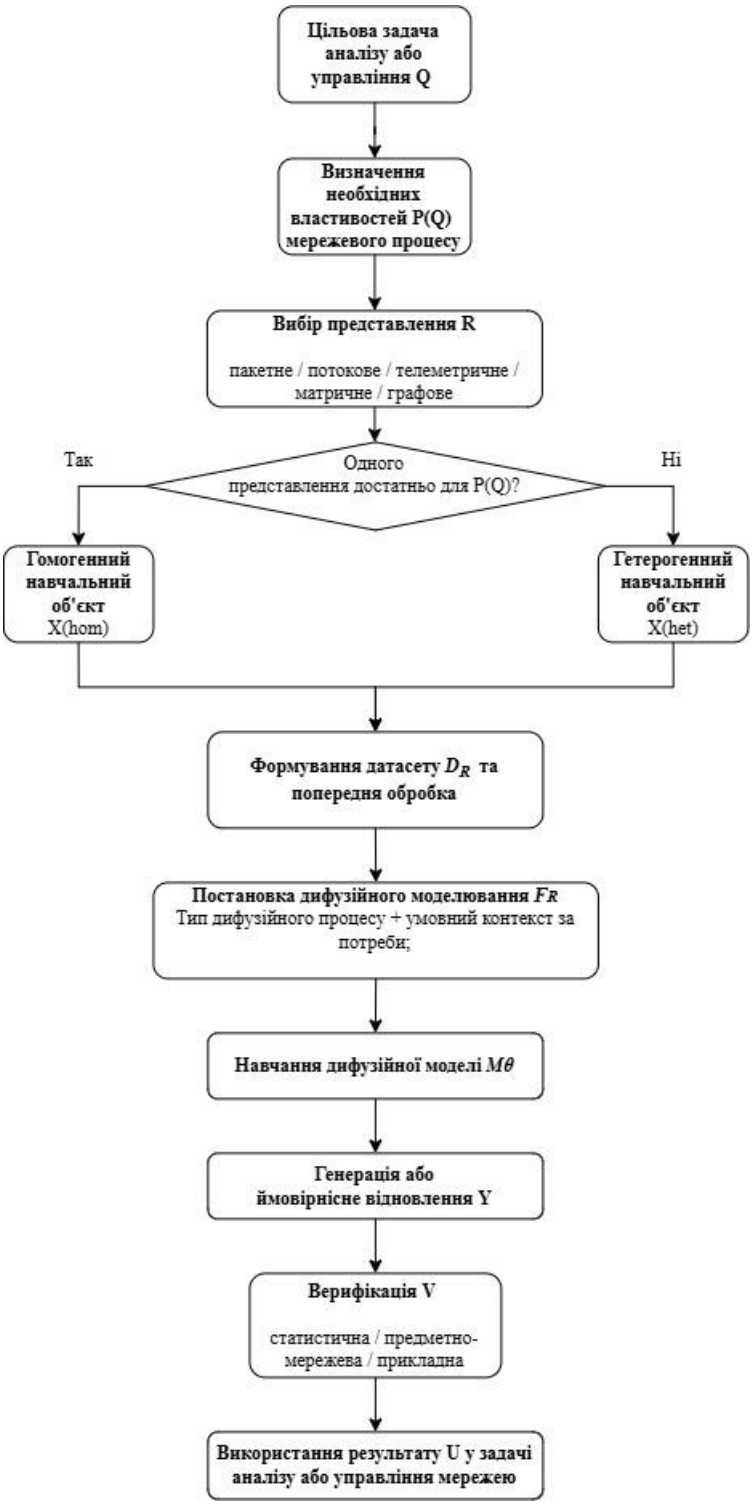


*Рис. 1.* **Узагальнена структура процесу формування, навчання, верифікації та використання дифузійної моделі для мережевих даних**

***Джерело*: розроблено авторами**

Інший варіант застосування послідовності (6) виникає у задачах Traffic Engineering, де для прийняття рішень можуть бути потрібні як просторово-часові характеристики трафікового попиту, так і інформація про топологію, маршрутизацію або доступні ресурси. Якщо цільова задача потребує генерації або відновлення лише матриці трафіку за фіксованої топології та маршрутизації, матричне представлення утворює гомогенний навчальний об'єкт, а топологічна й маршрутизаційна інформація використовується як умовний контекст *c*. Якщо ж до об'єкта генерації або відновлення одночасно включаються компоненти матричного та графового чи маршрутизаційного представлень, умова (2) для одного представлення є недостатньою, застосовується умова (4), а навчальний об'єкт формується як гетерогенний відповідно до (5). Такий приклад демонструє, що роль одних і тих самих мережевих даних визначається не їх природою, а способом участі у конкретній постановці дифузійного моделювання.

Верифікація V виконується відповідно до структури навчального об'єкта та призначення отриманого результату і включає статистичний, предметно-мережевий та прикладний рівні оцінювання, які детальніше розглянуто далі. Таким чином, вибір представлення мережевого трафіку розглядається не як ізольований етап підготовки даних, а як складова послідовного процесу, в якому вимоги цільової задачі визначають необхідні властивості даних, їх представлення, структуру навчального об'єкта, спосіб формування датасету та постановку дифузійного моделювання.

Наукова новизна роботи полягає у формалізованому представленні методологічного процесу вибору моделі представлення мережевого трафіку, формування та узгодження компонентів навчального об'єкта і постановки дифузійного моделювання відповідно до властивостей мережевого процесу, необхідних для розв'язання цільової задачі аналізу або управління комп'ютерною мережею. Встановлено та систематизовано відповідності між структурою сформованого навчального об'єкта, типами його компонентів, характером дифузійного процесу, використанням умовного контексту, змістом генерації або ймовірнісного відновлення та вимогами до верифікації результату. Запропоновано операційний критерій розмежування компонента навчального об'єкта та умовного контексту за характером їх участі в прямому та зворотному дифузійному процесі. У межах методологічного процесу розрізняються гомогенні навчальні об'єкти, сформовані на основі одного представлення, та гетерогенні навчальні об'єкти, що поєднують узгоджені компоненти декількох представлень. Для оцінювання результатів генерації та відновлення узагальнено статистичний, предметно-мережевий і прикладний рівні критеріїв.

**Дифузійне моделювання мережевих даних.** Після вибору представлення та формування навчального об'єкта його структура і типи компонентів визначають відповідну постановку дифузійного моделювання. Нехай $X_0$ позначає сформований навчальний об'єкт мережевих даних. Дифузійна модель навчається відтворювати розподіл саме таких об'єктів, тому зміст результату генерації або відновлення безпосередньо залежить від структури $X_0$.
Для неперервних числових компонентів базовий прямий дифузійний процес полягає у поступовому додаванні гаусівського шуму [1]. Нехай *s*, $s$=1,…,S, позначає номер кроку дифузійного процесу, де $S$ – загальна кількість кроків.

Тоді стан навчального об'єкта на кроці *s* можна подати як:

$$X_s = \sqrt{(\bar{\alpha}_s)}X_0 + \sqrt{(1 - \bar{\alpha}_s)}\varepsilon, \quad \varepsilon \sim N(0, I), \tag{7}$$

Де величина $\bar{\alpha}_s$ визначається співвідношенням:

$$\bar{\alpha}_s = \prod_{j=1}^{s}(1 - \beta_j), \tag{8}$$

де $\beta_j$ , $0 < \beta_j < 1$, визначає рівень шуму на відповідному дифузійному кроці. Відповідно до (7) та (8), зі збільшенням s внесок вихідного об'єкта $X_0$ зменшується, а внесок шумової складової зростає. Нейронна модель навчається реалізовувати зворотний процес, базова параметризація

якого задається розподілом переходу від $X_s$ до $X_{s-1}$ [1], [2]. В умовній постановці цей перехід додатково залежить від відомого контексту c.

Відповідний умовний розподіл можна подати як:

$$p\theta(X_{s-1} \mid X_s, c) = N(X_{s-1}; \mu\theta(X_s, s, c), \Sigma\theta(X_s, s, c)). \quad (9)$$

У співвідношенні (9) $\theta$ позначає параметри моделі, $\mu\theta$ – параметризоване середнє, $\Sigma\theta$ – параметри дисперсії або коваріації зворотного переходу, а $c$ – відома додаткова інформація, що використовується як умова генерації або відновлення. Відповідно до критерію, сформульованого вище, інформація $c$ не піддається відновленню як частина цільового об'єкта, а використовується для визначення умов зворотного дифузійного процесу. Прикладом такої постановки для ймовірнісного відновлення пропущених значень є модель CSDI [3].

Після навчання генерація нового зразка виконується шляхом послідовного застосування зворотних переходів від початкового шумового стану до об'єкта, що відповідає вивченому розподілу. Багатоетапність цього процесу збільшує обчислювальні витрати, тому для скорочення кількості кроків генерації можуть використовуватися прискорені схеми, зокрема метод DDIM [5].

Гаусівський дифузійний процес, представлений співвідношеннями (7) та (8), природно застосовується до неперервних характеристик, таких як обсяги трафіку, затримки, завантаження каналів або телеметричні показники після відповідної попередньої обробки. Водночас протоколи, типи повідомлень, стани, прапорці та інші категоріальні характеристики мають дискретну семантику. Просте числове кодування таких ознак не робить проміжні значення, які виникають унаслідок гаусівського зашумлення, предметно допустимими.

Для категоріальних компонентів можуть застосовуватися спеціалізовані дискретні дифузійні переходи, зокрема підходи, реалізовані в дискретних дифузійних моделях D3PM [8]. Якщо навчальний об'єкт одночасно містить неперервні та дискретні компоненти, доцільною є комбінована постановка, у якій механізм зашумлення та відновлення визначається типом відповідної ознаки. Можливість такого моделювання для змішаних даних демонструє модель TabDDPM [9].

Окремої уваги потребують категоріальні мережеві ознаки з великим простором можливих значень, зокрема IP-адреси та номери портів. Їх безпосереднє трактування як звичайних неперервних числових характеристик може створювати значення, що не мають коректної предметної інтерпретації. Тому спосіб їх кодування та дифузійного моделювання має враховувати семантику ознаки, структуру допустимого простору значень і вимоги конкретної задачі.

Умовне дифузійне моделювання відповідно до (9) дає змогу враховувати під час генерації або відновлення додаткову інформацію $c$. Для мережевих даних нею можуть бути режим навантаження, часові параметри, топологія, маршрутизація, характеристики каналів або доступна частина спостереження [3], [14]. Склад такого контексту визначається цільовою задачею і не залежить безпосередньо від того, чи сформовано гомогенний або гетерогенний навчальний об'єкт.

Для часових мережевих даних необхідно також розрізняти дифузійний крок $s$ і мережевий час $t$. Дифузійний крок $s$ у співвідношеннях (7)-(9) характеризує рівень зашумлення в межах генеративного процесу, тоді як мережевий час $t$ є складовою самого навчального об'єкта та визначає часову організацію пакетів, телеметричних вимірювань, матриць трафіку або графових станів. Отже, дифузійний процес змінює ступінь зашумлення об'єкта, але не замінює його часову структуру.

Таким чином, тип постановки дифузійного моделювання визначається насамперед природою компонентів сформованого навчального об'єкта: для неперервних характеристик застосовується неперервний процес, для дискретних – спеціалізовані дискретні переходи, а для їх поєднання – комбіноване моделювання. Використання умовного контексту

визначається окремо відповідно до задачі генерації або відновлення та характеризує відому інформацію, якою обумовлюється зворотний процес.

**Особливості дифузійного навчання для моделей представлення мережевого трафіку.** Пакетне, потокове, телеметричне, матричне та графове представлення відрізняються типами ознак, часовою і структурною організацією та характером внутрішніх залежностей [4]. Тому після вибору представлення відповідно до цільової задачі структура сформованого навчального об’єкта визначає тип дифузійного процесу, можливість використання умовного контексту та предметні вимоги до результату генерації або відновлення.

Пакетне представлення характеризується високою деталізацією та поєднанням неперервних, дискретних і бінарних характеристик. До числових компонентів можуть належати часові інтервали або розміри пакетів, тоді як протоколи, типи повідомлень, прапорці та службові поля мають дискретну семантику. Тому для такого навчального об’єкта доцільною є комбінована постановка дифузійного моделювання, у якій спосіб зашумлення залежить від типу відповідної ознаки [8], [9].

При цьому статистичної подібності згенерованих значень недостатньо: результат має зберігати допустимі співвідношення між полями, порядок пакетів і протокольну логіку. Такі обмеження можуть враховуватися на різних етапах генеративного процесу. Під час зворотного дифузійного процесу можуть застосовуватися механізми обмеженої або керованої генерації, за яких допустимість проміжного чи кінцевого стану враховується безпосередньо під час формування зразка. Іншим варіантом є проєкція отриманого результату на множину предметно допустимих станів. Після завершення генерації можуть використовуватися відсікання недопустимих зразків або постобробка на основі протокольних правил. Вибір конкретного механізму залежить від представлення, складності обмежень і цільової задачі. Необхідність спеціального врахування протокольної коректності демонструє модель NetDiffusion, орієнтована на генерацію пакетного мережевого трафіку з урахуванням протокольних обмежень [12].

Окрему проблему в пакетному та потоковому представленні становлять категоріальні ознаки з великою кількістю можливих значень, зокрема IP-адреси та номери портів. Безпосереднє подання таких характеристик як звичайних числових значень може створювати штучний порядок або відстань між категоріями, які не відповідають їх предметній семантиці. Крім того, великий простір можливих значень ускладнює пряме моделювання категоріального розподілу. Тому для таких ознак спосіб кодування має визначатися їх роллю в цільовій задачі: можливим є використання дискретного представлення, групування значень за функціональними категоріями, окреме моделювання структурно значущих полів або використання їх як умовного контексту, якщо самі значення не є цільовими компонентами генерації.

Потокове представлення агрегує властивості послідовності пакетів і також може містити числові та категоріальні компоненти. Дифузійна модель у такій постановці може формувати тривалість потоку, кількість пакетів і байтів, напрямок передачі, протокол та інші агреговані характеристики. Отримані значення мають залишатися логічно узгодженими, зокрема за співвідношеннями між тривалістю, кількістю пакетів і обсягом переданих даних. Для забезпечення такої узгодженості можуть застосовуватися обмеження під час генерації, перевірка допустимості сформованих комбінацій або їх корекція на етапі постобробки. Можливість дифузійної генерації потокових трас демонструє модель NetDiff, де формування потоків додатково обумовлюється сервісним контекстом [11].

Телеметричне представлення формує багатовимірні часові послідовності, значна частина компонентів яких є числовими. Після нормалізації та врахування семантики відповідних показників до них може застосовуватися неперервний дифузійний процес. Важливою для такого представлення є умовна постановка, за якої доступна частина часового ряду використовується для ймовірнісного відновлення пропущених значень [3], [7]. Попередня обробка при цьому має враховувати скидання та переповнення лічильників, нерівномірні

інтервали опитування, похідні показники та взаємозалежності між характеристиками обладнання і трафіку. Предметна коректність результату може перевірятися через допустимі діапазони показників, семантику монотонних або накопичувальних лічильників та узгодженість між пов'язаними часовими рядами.

Матричне представлення формує числовий просторово-часовий об'єкт, елементи якого характеризують трафіковий попит між парами джерело–призначення (origin–destination, OD). Для такого представлення дифузійна модель повинна відтворювати не лише розподіл окремих елементів, а й їх просторові та часові взаємозалежності. Умовний контекст може включати режим навантаження, часові характеристики, маршрутизацію або доступні мережеві ресурси залежно від цільової задачі.

Оскільки матриця трафіку описує OD-попит, а не безпосереднє завантаження каналів, маршрутизаційна та ресурсна узгодженість має враховуватися насамперед у задачах, де сформований попит використовується для аналізу каналів, маршрутизації або керування ресурсами. У таких випадках перевірка предметної допустимості може включати перетворення згенерованого OD-попиту відповідно до заданої маршрутизації та подальшу перевірку отриманого навантаження щодо ресурсних обмежень мережі. Застосування дифузійного підходу до аналізу та відновлення матриць трафіку за неповних спостережень демонструє підхід Diffusion-TM [13].

Графове представлення поєднує структурну інформацію про вузли та ребра з їх ознаками. Для задач управління комп'ютерними мережами доцільно розрізняти генерацію самої топології та генерацію або відновлення динамічних характеристик на відомому графі. У другому випадку дифузійний процес може застосовуватися до ознак вузлів і каналів, тоді як топологія відповідно до введеного вище критерію використовується як структурний умовний контекст. Якщо генерується сама графова структура, топологічні компоненти входять до навчального об'єкта та потребують дискретних механізмів для формування вузлів і ребер.

Предметні обмеження графового представлення можуть включати допустимість структури зв'язків, характеристики вузлів і каналів та їх узгодженість із топологією. Їх забезпечення також може виконуватися під час генерації шляхом обмеження простору допустимих графових переходів або після генерації шляхом перевірки й відсікання структур, що порушують задані вимоги. Можливість дифузійного моделювання графової структури демонструє модель DiGress [15], а застосування умовної графової дифузії до задач мережевого планування – модель GDPlan [16].

Таким чином, предметні обмеження можуть враховуватися на трьох основних етапах: безпосередньо під час зворотного дифузійного процесу через обмежену або керовану генерацію; після окремого кроку чи завершення генерації шляхом проєкції результату на допустиму множину; або після формування зразка через перевірку, відсікання чи корекцію відповідно до предметних правил. Ці способи не є взаємовиключними й можуть комбінуватися. У цій роботі вони розглядаються як загальні механізми забезпечення предметно-мережевої коректності, а вибір конкретної реалізації залежить від структури навчального об'єкта та архітектури дифузійної моделі.

Тип дифузійного процесу та використання умовного контексту при цьому залишаються різними характеристиками постановки. Тип процесу визначається насамперед природою компонентів навчального об'єкта, тоді як умовний контекст визначає додаткову інформацію, що враховується під час генерації або відновлення. Отже, одному представленню не відповідає єдиний алгоритм або єдиний спосіб дифузійного моделювання: конкретна постановка визначається структурою сформованого об'єкта, предметними обмеженнями та цільовою задачею.

Відповідність моделей представлення мережевих даних особливостям дифузійного моделювання узагальнено в Таблиці 1.

*Таблиця 1.* **Відповідність моделей представлення мережевих даних особливостям дифузійного моделювання**

| Представлення | Характер даних і тип дифузійного процесу | Умовний контекст та предметні обмеження | Результат |
|---|---|---|---|
| Пакетне | Неперервні, дискретні та бінарні ознаки; переважно комбінований процес | За потреби – параметри сценарію або стану обміну; протокольна логіка, залежності полів, порядок пакетів | Пакетна послідовність |
| Потокове | Числові та категоріальні ознаки; неперервний або комбінований процес | За потреби – сервісний або інший контекст; узгодженість тривалості, кількості пакетів, байтів та інших агрегованих характеристик | Послідовність потоків |
| Телеметричне | Переважно числові часові ряди; неперервний процес | Умовна постановка для відновлення або керованої генерації; доступні спостереження, час, семантика лічильників | Генерація або відновлення телеметрії |
| Матриця трафіку | Числові просторово-часові дані; переважно неперервний процес | Умовна постановка за часткових спостережень або заданого мережевого контексту; часові, просторові та, за потреби, маршрутизаційні й ресурсні залежності | Матриця або послідовність матриць попиту |
| Графове | Ознаки вузлів і ребер та/або структура; неперервний, дискретний чи комбінований процес залежно від генерованих компонентів | Топологія може використовуватися як структурна умова; узгодженість ознак вузлів і ребер зі структурою графа | Графовий стан або його компоненти |

***Джерело*: розроблено авторами**

Кожне з розглянутих представлень може самостійно утворювати гомогенний навчальний об’єкт. Гетерогенний об’єкт формується, якщо цільова задача потребує спільного генерування або відновлення компонентів декількох представлень, наприклад телеметричних показників разом із графовими характеристиками вузлів і каналів або матриці трафіку разом із іншими компонентами мережевого стану. Відповідно до (5), формування такого об’єкта потребує узгодження часових масштабів, частот дискретизації, способів кодування та структурних ідентифікаторів компонентів. Якщо ж додаткові топологічні, маршрутизаційні або ресурсні дані залишаються відомими та лише обумовлюють генерацію цільового представлення, вони належать до умовного контексту c, а не утворюють додатковий компонент гетерогенного об’єкта.

Отже, структура представлення визначає не лише форму даних, а й предметні залежності, які повинні зберігатися після генерації або відновлення. Для пакетних і потокових об’єктів суттєвими є змішані типи ознак та логічні залежності, для телеметричних і матричних – часові та просторові взаємозв’язки, а для графових – узгодженість динамічних характеристик із топологічною структурою мережі. Забезпечення цих властивостей може вимагати не лише відповідного дифузійного процесу, а й спеціальних механізмів обмеження, перевірки або корекції результату.

**Відновлення неповних і недостовірних даних та критерії якості**. Одним із практично важливих застосувань умовного дифузійного моделювання є ймовірнісне відновлення

відсутніх компонентів за доступною частиною спостереження. Нехай $B$ — бінарна маска тієї самої розмірності, що й навчальний об'єкт $X$, де одиниця відповідає доступному компоненту.

Тоді спостережувана частина об'єкта визначається як:

$$X^{obs} = B \odot X, \tag{10}$$

а відсутня частина – як:

$$X^{miss} = (1 - B) \odot X, \tag{11}$$

де $\odot$ позначає покомпонентне множення. Відповідно до (10) і (11), маска B визначає, яка частина мережевого об'єкта безпосередньо доступна моделі як спостереження, а яка підлягає відновленню.

Умовне ймовірнісне відновлення відсутніх компонентів можна подати як

$$\tilde{X}^{miss} \sim p\theta(X^{miss} \mid X^{obs}, B, c), \tag{12}$$

де $\tilde{X}^{miss}$ – відновлений варіант відсутньої частини об'єкта, а $c$ – додатковий умовний контекст. Співвідношення (12) означає, що дифузійна модель формує вибірку з умовного розподілу можливих значень невідомої частини за доступними спостереженнями, маскою пропусків і додатковим контекстом [3]. Тому результат такого відновлення не є детермінованою оцінкою фактично втрачених значень, а представляє один із можливих станів, узгоджених із доступною інформацією та розподілом, засвоєним моделлю.

Зміст маски B і контексту c залежить від представлення мережевих даних. Для телеметрії маска може позначати окремі пропущені показники або часові інтервали, для матриці трафіку – невідомі елементи OD-попиту, а для графового представлення – відсутні характеристики вузлів чи каналів. Застосування дифузійного підходу до відновлення неповних матриць трафіку, зокрема, розглядається в Diffusion-TM [13]. Для пакетних і потокових об'єктів відновлені компоненти додатково повинні зберігати структурні, логічні та протокольні залежності. Отже, відновлення окремих значень за (12) саме по собі не гарантує коректності мережевого об'єкта в цілому.

Особливе значення для мережевих даних має механізм виникнення пропусків. У найпростішому випадку відсутність окремих значень може бути випадковою і не залежати безпосередньо від стану мережі. Однак на практиці зустрічаються також групові або часові пропуски, спричинені недоступністю джерела даних, втратою послідовності повідомлень або порушенням процесу збору телеметрії. Крім того, пропуски можуть бути пов'язані безпосередньо з досліджуваним станом мережі. Наприклад, перевантаження, відмова обладнання, втрата зв'язності або нестача ресурсів системи моніторингу можуть одночасно змінювати мережеві показники та збільшувати ймовірність втрати телеметричних повідомлень чи фрагментів трас.

У такому випадку маска B сама містить інформацію про стан мережі, а припущення про незалежність механізму пропусків від даних може призводити до зміщення результатів відновлення. Тому під час формування навчальних і контрольних вибірок доцільно розрізняти принаймні випадкові точкові пропуски, послідовні або блокові пропуски та пропуски, пов'язані з мережевими подіями. Для останнього випадку інформація про причину або умови втрати даних, якщо вона доступна, може включатися до умовного контексту c. Під час верифікації доцільно окремо оцінювати якість відновлення для різних механізмів пропусків, оскільки результат, отриманий за штучно випадково прихованих значень, може не відображати якість роботи моделі в умовах реальних перевантажень або відмов.

При цьому необхідно розрізняти неповні та недостовірні спостереження. Для пропущеного компонента його положення відоме завдяки масці B, тоді як помилкове або спотворене значення може формально залишатися частиною $X^{obs}$. Дифузійна модель сама по собі не забезпечує автоматичного виявлення таких значень. Тому робота з недостовірними спостереженнями потребує попередньої ідентифікації ненадійних компонентів або

оцінювання рівня довіри до них, після чого відповідна інформація може враховуватися через маску, умовний контекст або інший механізм постановки задачі.

Результат відновлення відповідно до (12) не слід інтерпретувати як гарантоване відтворення фактично втраченого мережевого стану. Одна й та сама доступна частина спостереження може відповідати декільком можливим станам мережі. Тому дифузійна модель може формувати декілька умовно правдоподібних варіантів, а невизначеність між ними є складовою результату ймовірнісного відновлення. Для задач управління мережею це принципово, оскільки використання одного реконструйованого варіанта без урахування його невизначеності може призводити до надмірно впевненого подальшого рішення.

Верифікацію результатів генерації та відновлення доцільно виконувати на трьох взаємодоповнювальних рівнях: статистичному, предметно-мережевому та прикладному. Конкретний набір метрик визначається структурою навчального об’єкта, характером результату та цільовою задачею; тому жоден окремий показник не може розглядатися як універсальна оцінка якості мережевої генерації.

На статистичному рівні оцінюється відповідність розподілів і залежностей між реальними та синтетичними або відновленими даними. Для цього можуть використовуватися дивергенція Дженсена-Шеннона (JSD), відстань Вассерштейна, максимальна середня розбіжність (Maximum Mean Discrepancy, MMD), а також класифікаційний тест двох вибірок (two-sample classifier test), у якому оцінюється можливість відрізнити реальні об’єкти від згенерованих. Для часових даних додатково доцільно порівнювати автокореляційні характеристики, сезонність, розподіли інтервалів та міжознакові залежності.

Для синтетичної генерації статистичний рівень має також враховувати різноманітність і можливе надмірне відтворення навчальних зразків. Для цього можуть використовуватися відстань до найближчого навчального об’єкта (distance to closest record, DCR) і відношення відстаней до найближчих сусідів (nearest-neighbour distance ratio). Такі показники не характеризують предметну коректність безпосередньо, але дають змогу додатково оцінити, чи не складається синтетична вибірка переважно з близьких копій навчальних прикладів.

Для задач відновлення, коли еталонні значення відомі, наприклад у разі штучного маскування частини контрольних даних, можуть використовуватися MAE, RMSE або інші показники похибки відповідно до типу ознак. Водночас оцінювання лише за випадково прихованими компонентами є недостатнім, якщо в реальній мережі пропуски мають блоковий або залежний від мережевого стану характер. Тому контрольні сценарії доцільно формувати відповідно до декількох механізмів пропусків.

Предметно-мережевий рівень визначає, чи зберігає сформований результат обмеження та залежності відповідного представлення. Для пакетних даних можуть оцінюватися частка протокольно валідних пакетів, частота порушень залежностей між полями та коректність послідовності пакетів. Для потокового представлення перевіряється узгодженість тривалості, кількості пакетів, обсягу байтів та інших агрегованих характеристик. Для телеметрії можуть оцінюватися порушення допустимих діапазонів, семантики накопичувальних лічильників, коректність скидань або переповнень та узгодженість взаємопов’язаних показників.

Для матриць трафіку предметна перевірка охоплює збереження просторової та часової структури OD-попиту. Якщо цільова задача пов’язана з Traffic Engineering, маршрутизацією або керуванням ресурсами, додатково може оцінюватися узгодженість навантаження, отриманого з матриці трафіку за заданою маршрутизацією, з пропускною здатністю та іншими ресурсними обмеженнями мережі. Для графових об’єктів відповідні критерії характеризують допустимість топологічної структури та узгодженість ознак вузлів і каналів із графом.

Прикладний рівень визначає, наскільки сформовані дані придатні до використання в цільовій задачі. Для оцінювання синтетичних наборів можуть застосовуватися схеми Train on Synthetic, Test on Real (TSTR) та Train on Real, Test on Synthetic (TRTS), які порівнюють корисність синтетичних і реальних даних для навчання та перевірки прикладної моделі. Конкретні показники при цьому визначаються задачею: для виявлення аномалій це можуть

бути F1-score, precision, recall або AUROC, для прогнозування та оцінювання навантаження – MAE або RMSE, а для задач маршрутизації чи Traffic Engineering – показники використання ресурсів, перевантаження каналів або інші цільові характеристики алгоритму управління.

Відповідність рівнів верифікації конкретним групам показників доцільно узагальнити в Табл. 2.

*Таблиця 2.* **Рівні оцінювання результатів дифузійної генерації та відновлення мережевих даних**

| Рівень оцінювання | Зміст | Приклади |
|---|---|---|
| Статистичний | Відповідність статистичних розподілів, часових і міжознакових залежностей, різноманітність синтетичних зразків та відсутність надмірного відтворення навчальних даних | JSD; Wasserstein distance; MMD; two-sample classifier test; кореляції та автокореляції; MAE/RMSE для контрольного відновлення; DCR; nearest-neighbour distance ratio |
| Предметно-мережевий | Дотримання обмежень і залежностей конкретного представлення | Частка протокольно валідних пакетів; частота порушень залежностей полів; послідовнісна коректність; узгодженість duration–packets–bytes для потоків; семантична й часова коректність телеметричних показників і лічильників; просторово-часова узгодженість OD-попиту; за потреби – відповідність маршрутизаційним, ресурсним і топологічним обмеженням |
| Прикладний | Придатність синтетичних або відновлених даних до цільової задачі | TSTR; TRTS; F1-score, precision, recall, AUROC для виявлення аномалій; MAE/RMSE для прогнозування або оцінювання навантаження; показники ефективності маршрутизації та Traffic Engineering |

***Джерело*: розроблено авторами**

Наведені рівні не є взаємозамінними. Низька похибка відновлення на контрольних компонентах не гарантує предметно-мережевої допустимості результату. Аналогічно, статистично близький до реальних даних синтетичний об’єкт може порушувати протокольні, часові, маршрутизаційні або топологічні залежності, а предметно коректний зразок не обов’язково забезпечує приріст якості в прикладній задачі. Тому висновок про якість генерації або відновлення доцільно формувати лише на основі сукупності показників, вибраних відповідно до представлення та призначення результату.

Таким чином, верифікація V у послідовності (6) має включати не універсальну метрику, а набір взаємодоповнювальних критеріїв. Для задач відновлення цей набір додатково має враховувати механізм виникнення пропусків, а для синтетичної генерації — різноманітність, відсутність надмірного копіювання навчальних зразків, предметну коректність та прикладну придатність отриманих даних.

Обговорення результатів та обмеження. Проведений аналіз показує, що вибір представлення мережевого трафіку доцільно розглядати як складову послідовного процесу побудови та використання дифузійної моделі, узагальненого послідовністю (6). Вимоги цільової задачі визначають необхідні властивості мережевих даних, на основі яких обирається одне або декілька представлень, формується навчальний об’єкт і датасет,

визначається постановка дифузійного моделювання та виконується верифікація результату. Таким чином, представлення даних впливає не лише на форму вхідної інформації, а й на зміст об’єкта генерації або відновлення та критерії його коректності.

Встановлені відповідності не утворюють однозначного зв’язку «одне представлення – один тип дифузійної моделі». Конкретна постановка залежить від складу навчального об’єкта, типів його компонентів і цільової задачі. Неперервні, дискретні та змішані компоненти потребують відповідних механізмів дифузійного процесу, тоді як умовний контекст визначає додаткову відому інформацію, що використовується під час генерації або відновлення.

Розрізнення гомогенних і гетерогенних навчальних об’єктів дає змогу формалізувати випадки, коли властивостей одного представлення достатньо для цільової задачі або необхідне поєднання декількох представлень. Формування гетерогенного об’єкта відповідно до (5) передбачає узгодження часових масштабів, частот дискретизації, способів кодування та структурних ідентифікаторів його компонентів. При цьому додаткова інформація, яка залишається відомою та лише обумовлює зворотний дифузійний процес, належить до умовного контексту c і сама по собі не робить навчальний об’єкт гетерогенним.

Структура представлення визначає також предметні залежності, які повинні зберігатися в результаті. Для пакетних даних такими є протокольна та послідовнісна коректність, для потокових – узгодженість агрегованих характеристик, для телеметрії – семантика показників і часові залежності, для матриць трафіку – просторова та часова структура OD-попиту, а для графових об’єктів — відповідність характеристик топологічній структурі. Такі обмеження можуть враховуватися безпосередньо під час генерації або через перевірку, проєкцію, відсікання чи постобробку результату. Тому статистична подібність до реальних даних є необхідною, але недостатньою характеристикою якості.

Окреме практичне значення має ймовірнісне відновлення неповних спостережень відповідно до (12). Воно дозволяє формувати правдоподібні варіанти невідомих компонентів, але не гарантує відтворення фактично втраченого стану. Крім того, у мережевих даних пропуски можуть бути не лише випадковими, а й пов’язаними з перевантаженнями, відмовами або порушеннями процесу моніторингу. У таких випадках механізм виникнення пропусків необхідно враховувати під час формування контрольних сценаріїв та інтерпретації результатів відновлення. Недостовірні значення додатково потребують попередньої ідентифікації або оцінювання рівня довіри до них.

Запропонована формалізація узагальнює взаємозв’язок між вимогами цільової задачі, вибором представлення мережевих даних, формуванням гомогенного або гетерогенного навчального об’єкта, постановкою дифузійного моделювання та верифікацією результату. Наведені наскрізні приклади конкретизують застосування послідовності (6) для типових мережевих задач і показують, як зміна цільових вимог впливає на склад навчального об’єкта та роль умовного контексту. Межі застосовності запропонованого підходу пов’язані з тим, що конкретний спосіб узгодження компонентів, архітектура дифузійної моделі та набір критеріїв верифікації мають визначатися з урахуванням структури вихідних даних і особливостей цільової задачі. Експериментальне оцінювання впливу цих факторів становить напрям подальших досліджень.

**Висновки.** У роботі формалізовано методологічний процес вибору представлення мережевого трафіку, формування навчального об’єкта, підготовки датасету, постановки дифузійного моделювання, генерації або відновлення, верифікації та подальшого використання отриманих даних відповідно до цільової задачі аналізу або управління комп’ютерною мережею. Вибір представлення запропоновано здійснювати на основі множини властивостей мережевого процесу, необхідних для розв’язання конкретної задачі. Для пакетного, потокового, телеметричного, матричного та графового представлень конкретизовано властивості, що визначають їх придатність до формування відповідного навчального об’єкта.

Якщо необхідні властивості забезпечуються одним представленням, формується гомогенний навчальний об’єкт; якщо їх забезпечення потребує компонентів декількох представлень – гетерогенний. Для гетерогенних об’єктів конкретизовано необхідність попереднього узгодження часових масштабів, частот дискретизації, способів нормалізації та кодування, а також структурних ідентифікаторів компонентів. Встановлено критерій розмежування складової навчального об’єкта та умовного контексту: компонент, що піддається дифузійному перетворенню і має бути згенерований або відновлений, належить до навчального об’єкта, тоді як відома інформація, що лише обумовлює зворотний процес, розглядається як умовний контекст. Таким чином, використання умовного контексту саме по собі не визначає гетерогенність навчального об’єкта.

Для розглянутих представлень встановлено відповідності між структурою навчального об’єкта, типами його компонентів, вибором неперервного, дискретного або комбінованого дифузійного процесу та предметними вимогами до результату. Показано, що статистичної подібності з реальними даними недостатньо для підтвердження коректності генерації. Залежно від представлення необхідно додатково забезпечувати протокольну та послідовнісну коректність пакетних даних, узгодженість агрегованих характеристик потоків, семантику телеметричних показників і лічильників, просторово-часову структуру OD-попиту та відповідність графових характеристик топології. Такі обмеження можуть враховуватися під час генеративного процесу або шляхом подальшої проєкції, перевірки, відсікання чи корекції сформованих результатів.

Для умов неповних спостережень уточнено постановку ймовірнісного відновлення відсутніх компонентів за доступною частиною даних, маскою спостережень і додатковим контекстом. Показано необхідність врахування механізму виникнення пропусків, оскільки в мережевих даних вони можуть бути випадковими, блоковими або пов’язаними з перевантаженнями, відмовами та іншими змінами мережевого стану. Результат відновлення слід трактувати як умовно правдоподібний варіант невідомої частини мережевого стану, а не як гарантоване відтворення фактично втрачених значень. Недостовірні спостереження, на відміну від відомих пропусків, потребують попередньої ідентифікації ненадійних компонентів або оцінювання рівня довіри до них.

Верифікацію результатів генерації та відновлення запропоновано здійснювати на статистичному, предметно-мережевому та прикладному рівнях із вибором конкретних показників відповідно до структури представлення та цільової задачі. Статистичні характеристики мають доповнюватися перевіркою предметних залежностей і обмежень, а також оцінюванням придатності отриманих даних до подальших задач аналізу або управління. Наведені наскрізні приклади відновлення мережевої телеметрії та використання матриць трафіку в задачах Traffic Engineering конкретизують застосування запропонованого процесу від визначення вимог цільової задачі до формування навчального об’єкта та верифікації результату.

Отримані результати формують методологічну основу для узгодженого вибору представлення мережевих даних і постановки дифузійного моделювання відповідно до задач аналізу та управління комп’ютерними мережами. Подальші дослідження доцільно спрямувати на експериментальне порівняння гомогенних і гетерогенних навчальних об’єктів, дослідження способів узгодження їх компонентів та оцінювання впливу структури навчальних даних і механізмів виникнення пропусків на якість генерації, відновлення й використання отриманих даних у конкретних мережевих задачах.

**Конфлікт інтересів:** автори заявляють про відсутність конфлікту інтересів

**Фінансування:** дослідження виконано без фінансової підтримки

**Доступність даних:** окремі набори даних у межах цього дослідження не створювалися та не аналізувалися; усі матеріали, необхідні для відтворення представленого методологічного аналізу, наведені в тексті статті та процитованих джерелах

**Використання штучного інтелекту:** Під час підготовки та редагування рукопису використовувався інструмент OpenAI ChatGPT (модель GPT-5.6 Sol) для мовно-стилістичного редагування, структурного аналізу тексту, перевірки логічної узгодженості та формулювання редакційних пропозицій. Усі запропоновані зміни, наукові положення, математичні формалізації, джерела та висновки були перевірені та остаточно затверджені авторами. Відповідальність за зміст роботи повністю несуть автори

## СПИСОК ЛІТЕРАТУРИ

1. Ho, J., Jain, A. & Abbeel, P. “Denoising diffusion probabilistic models”. *Advances in Neural Information Processing Systems*. 2020; 33: 6840–6851. DOI: https://doi.org/10.5555/3495724.3496298.

2. Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S. & Poole, B. “Score-based generative modeling through stochastic differential equations”. *International Conference on Learning Representations (ICLR). arXiv*. 2021. DOI: https://doi.org/10.48550/arXiv.2011.13456.

3. Tashiro, Y., Song, J., Song, Y. & Ermon, S. “CSDI: Conditional score-based diffusion models for probabilistic time series imputation”. *Advances in Neural Information Processing Systems*. 2021; 34: 24804–24816. DOI: https://doi.org/10.5555/3540261.3542161.

4. Науменко, Р., Нестеренко, С. та Буюклі, В. “Моделі представлення мережевого трафіку в задачах аналізу та управління комп’ютерними мережами”. *Věda a perspektivy*. 2026; 7 (62): 462–475. DOI: https://doi.org/10.52058/2695-1592-2026-7(62)-462-475.

5. Song, J., Meng, C. & Ermon, S. “Denoising diffusion implicit models”. *International Conference on Learning Representations (ICLR). arXiv*. 2021. DOI: https://doi.org/10.48550/arXiv.2010.02502.

6. Rasul, K., Seward, C., Schuster, I. & Vollgraf, R. “Autoregressive denoising diffusion models for multivariate probabilistic time series forecasting”. *Proceedings of the 38th International Conference on Machine Learning. Proceedings of Machine Learning Research*. 2021; 139: 8857–8868. DOI: https://doi.org/10.48550/arXiv.2101.12072.

7. Yuan, X. & Qiao, Y. “Diffusion-TS: Interpretable diffusion for general time series generation”. *International Conference on Learning Representations (ICLR).* 2024. DOI: https://doi.org/10.48550/arXiv.2403.01742.

8. Austin, J., Johnson, D. D., Ho, J., Tarlow, D. & van den Berg, R. “Structured denoising diffusion models in discrete state-spaces”. *Advances in Neural Information Processing Systems*. 2021; 34: 17981–17993. DOI: https://doi.org/10.5555/3540261.3541637.

9. Kotelnikov, A., Baranchuk, D., Rubachev, I. & Babenko, A. “TabDDPM: Modelling tabular data with diffusion models”. *Proceedings of the 40th International Conference on Machine Learning. Proceedings of Machine Learning Research*. 2023; 202: 17564–17579. DOI: https://doi.org/10.5555/3618408.3619133.

10. Sivaroopan, N., Bandara, D., Madarasingha, C., Jourjon, G., Jayasumana, A. P. & Thilakarathna, K. “NetDiffus: Network traffic generation by diffusion models through time-series imaging”. *Computer Networks*. 2024; 251: 110616. DOI: https://doi.org/10.1016/j.comnet.2024.110616.

11. Zhang, S., Li, T., Jin, D. & Li, Y. “NetDiff: A service-guided hierarchical diffusion model for network flow trace generation”. *Proceedings of the ACM on Networking*. 2024; 2 (CoNEXT3): 16. DOI: https://doi.org/10.1145/3676870.

12. Jiang, X., Liu, S., Gember-Jacobson, A., Bhagoji, A. N., Schmitt, P., Bronzino, F. & Feamster, N. “NetDiffusion: Network data augmentation through protocol-constrained traffic generation”. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*. 2024; 8 (1): 11, https://www.scopus.com/pages/publications/85185882002.
DOI: https://doi.org/10.1145/3639037.

13. Yuan, X., Qiao, Y., Wei, Z., Zhang, Z., Li, M., Zhao, P., Hu, R. & Li, W. “Diffusion models meet network management: Improving traffic matrix analysis with diffusion-based approach”. *IEEE Transactions on Network and Service Management*. 2025; 22 (2): 1259–1275. D OI: https://doi.org/10.1109/TNSM.2025.3527442.

14. Chai, H., Qi, X. & Li, Y. “Spatio-Temporal Knowledge Driven Diffusion Model for Mobile Traffic Generation”. *IEEE Transactions on Mobile Computing*. 2025; 24 (6): 4939–4956. DOI: https://doi.org/10.1109/TMC.2025.3527966.

15. Vignac, C., Krawczuk, I., Siraudin, A., Wang, B., Cevher, V. & Frossard, P. “DiGress: Discrete denoising diffusion for graph generation”. *International Conference on Learning Representations (ICLR).* 2023. DOI: https://doi.org/10.48550/arXiv.2209.14734.

16. Kan, N., Yan, S., Zou, J., Dai, W., Gao, X., Li, C. & Xiong, H. “GDPlan: Generative network planning via graph diffusion model”. *IEEE Transactions on Networking*. 2025; 33 (4): 1422–1437. DOI: https://doi.org/10.1109/TON.2025.3535518.

Отримано 03.08.2026
Отримано після перегляду 28.08.2026
Прийнято 04.09.2026

**DOI: https://doi.org/10.15276/ict.03.2026.30**
**UDC 004.8:004.7**

# Analysis of network traffic representation models for training diffusion models in computer network management tasks

**Roman I. Naumenko[1)]**
ORCID: https://orcid.org/0009-0003-7612-1773; 9530675@stud.op.edu.ua
**Serhii A. Nesterenko[1)]**
ORCID: https://orcid.org/0000-0002-3757-6594; san@op.edu.ua. Scopus Author Id: 58596482600
[1)]Odesa Polytechnic National University, 1, Shevchenko Ave. Odesa, 65044,Ukraine

## ABSTRACT

The paper examines the use of network traffic representation models as training objects for diffusion generative models in computer network analysis and management tasks. The relevance of the study is determined by the fact that real network data are often characterized by missing values, irregular measurements, distortions, storage limitations, and insufficient representation of particular network operating conditions. Under such circumstances, diffusion models can be used both to generate synthetic samples and, in a conditional setting, to probabilistically reconstruct missing components; however, the way they are applied depends on the structure of the training object. Based on packet-level, flow-level, telemetry, traffic-matrix, and graph-based representations of network data, the correspondence between feature types, temporal and structural organization of the data, and the choice of continuous, discrete, or combined diffusion modeling is analyzed. The process of selecting a representation according to the properties of the network process required to solve a target task is formalized, including the subsequent formation of the training object and dataset, specification of the diffusion modeling setting, model training, verification, and use of the obtained result. Domain-specific constraints of the corresponding representations are analyzed, including protocol and sequential correctness of packet data, consistency of aggregated flow characteristics, semantics of telemetry indicators and counters, spatiotemporal dependencies of traffic matrices and, depending on the target task, their consistency with routing and resource context, as well as the consistency of dynamic characteristics with the topological structure of graph-based representations. A distinction is proposed between homogeneous training objects formed within a single representation and heterogeneous training objects combining components of several network data representations. It is shown that the heterogeneity of a training object is not equivalent to the heterogeneity of feature types and should be considered separately from the use of additional known information as a condition for generation or reconstruction. A probabilistic formulation for reconstructing incomplete network observations from the known part of the data and additional context is formalized. It is substantiated that the reconstructed result should be interpreted as a conditionally plausible network state rather than as a guaranteed recovery of the actual missing values. Three complementary levels of criteria are generalized for verifying the results of diffusion-based generation and reconstruction: statistical, network-domain, and application-level. The obtained results can serve as a methodological basis for selecting a network data representation, forming a training object, and specifying a diffusion modeling setting according to the target computer network analysis or management task.

**Keywords:** Network traffic; diffusion models; network data representation; network traffic generation; missing data reconstruction; homogeneous training objects; heterogeneous training objects; computer network management

## ІНФОРМАЦІЯ ПРО АВТОРІВ

**Науменко Роман Іванович** - аспірант кафедри Комп’ютерних інтелектуальних систем та мереж. Національний університет «Одеська політехніка», пр. Шевченка, 1. Одеса, 65044,Україна
***Галузь досліджень:*** методи та системи штучного інтелекту

**Roman I. Naumenko** - PhD student, Department of Computer Intelligent Systems and Networks. Odesa Polytechnic National University, 1, Shevchenko Ave. Odesa, 65044, Ukraine
ORCID: https://orcid.org/0009-0003-7612-1773; 9530675@stud.op.edu.ua
***Research field:*** artificial intelligence methods and systems;

**Нестеренко Сергій Анатолійович** - доктор технічних наук, професор кафедри Комп’ютерних інтелектуальних систем та мереж. Національний університет «Одеська політехніка», пр. Шевченка, 1. Одеса, 65044, Україна
***Галузь дослідження:*** комп'ютерні мережі та інформаційні технології

**Serhii A. Nesterenko** - Doctor of Engineering Science, Professor, Department of Computer Intelligent Systems and Networks. Odesa Polytechnic National University, 1, Shevchenko Ave. Odesa, 65044, Ukraine
ORCID: https://orcid.org/0000-0002-3757-6594; san@op.edu.ua. Scopus Author Id: 58596482600
***Research field:*** computer networks and information technologies