МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
Національний авіаційний університет

ДОЛГИХ Сергій Миколайович

УДК 004.7:004.032.26

ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ РОЗПІЗНАВАННЯ МЕРЕЖЕВИХ ДАНИХ ІНТЕРНЕТ НА ОСНОВІ ГЕНЕРАТИВНИХ НЕЙРОМЕРЕЖЕВИХ МОДЕЛЕЙ

Спеціальність 05.13.06 – «Інформаційні технології»

Автореферат

дисертації на здобуття наукового ступеня
кандидата технічних наук

Київ 2023

Дисертацією є рукопис.

Робота виконана на кафедрі прикладної математики в Національному авіаційному університеті Міністерства освіти і науки України.

Науковий керівник: доктор технічних наук, професор
Приставка Пилип Олександрович,
Національний авіаційний університет,
завідувач кафедри прикладної математики.

Офіційні опоненти: доктор технічних наук, професор
Терейковський Ігор Анатолійович,
Національний технічний університет України «Київський політехнічний інститут імені Ігоря Сікорського», професор кафедри системного програмування і спеціалізованих комп'ютерних систем;

кандидат технічних наук, доцент
Мирутенко Лариса Вікторівна,
Київський національний університет імені Тараса Шевченка, доцент кафедри кібербезпеки та захисту інформації.

Захист відбудеться 01 червня 2023 року о $14^{\underline{00}}$ на засіданні спеціалізованої вченої ради Д 26.062.01 при Національному авіаційному університеті за адресою: 03058, м. Київ, пр. Любомира Гузара, 1, корпус 6, аудиторія 102.

З дисертацією можна ознайомитись у Науково-технічній бібліотеці Національного авіаційного університету за адресою: 03058, м. Київ, пр. Любомира Гузара, 1.

Автореферат розісланий 26 квітня 2023 року.

Учений секретар
спеціалізованої вченої ради
кандидат технічних наук Тетяна ОХРІМЕНКО

ЗАГАЛЬНА ХАРАКТЕРИСТИКА РОБОТИ

Актуальність. Методи та технології машинного навчання досягли суттєвого прогресу за останні роки, знайшовши успішне застосування в широкому діапазоні галузей і додатків, який постійно розширюється. У кількох галузях, таких як: класифікація зображень; деякі сфери медичної діагностики; деякі творчі заняття та ігри, такі як шахи та Ґо; та багато інших, здібності машинних систем приймати розумні рішення наближаються, або в деяких випадках, перевершують можливості середньостатистичної людини і навіть людини-експерта. Ґрунтуючись на цих досягненнях, з'являється значна впевненість у тому, що як галузь застосування, так і продуктивність систем штучного інтелекту продовжуватимуть зростати та розширюватися.

Одночасно, попри ці потужні досягнення стає ясно що можливості, які використовуються у традиційних методах штучного інтелекту, мають низку внутрішніх обмежень. По-перше, такі програми у складних задачах застосування систем машинного інтелекту вимагають величезних обсягів попереднього знання галузі в будь-якій з відомих форм, наприклад, масивні марковані навчальні набори даних із заздалегідь відомими структурами категорій. Крім того, машинні системи навчені такими методами можуть набувати лише статичні, фіксовані в часі знання і не можуть самі по собі вчитися розпізнавати нові концепції або змінювати структури вже вивчених концепцій що у багатьох сучасних системах машинного навчання вимагатиме повного перенавчання. У разі динамічних, мінливих середовищ застосування та даних цей такий процес не є дуже ефективним, зокрема у середовищах де недоступні значні обсяги навчальних даних маркованих відомими категоріями. Такі методи також не цілком відповідають процесам навчання біологічних систем, які є гнучкими, часто спонтанними, легко адаптованими та керованими середовищем і вимагають значно менших обсягів відомих даних для успішного навчання.

У вираженій формі ці питання та виклики виникають у задачах аналізу та класифікації даних Інтернет, таких як визначення типу трафіку, класифікація приналежності до додатків-джерел та аналогічних, що набувають значної актуальності в комерційних і соціальних додатках, як і в питаннях захисту, аналізу даних та інформаційної безпеки. Серйозною проблемою залишається наявність значних сучасних наборів навчальних даних для розробки систем розпізнавання трафіку та додатків до Інтернету; крім того, як показали результати декількох досліджень, навіть за наявності навчальних даних успіх навчання не може бути гарантований, тому що параметри трафіку в різних мережах можуть мати значну варіацію і успішне навчання на даних отриманих з «базової» мережі може не забезпечити успіху з реальними даними з іншої. Успішне застосування методів навчання систем машинного інтелекту у таких задачах та галузях повинні враховувати особливості розподілу параметрів у конкретних локальних мережах, де значні навчальні набори можуть вимагати значного часу та зусиль для підготовки, або бути відсутніми взагалі.

У результаті таких обмежень галузі виникає задача розробки систем штучного інтелекту здатних навчатися безпосередньо з немаркованими даними, не спираючись на значні системи заздалегідь відомих концептуальних структур; використовувати локальні, натуральні особливості та структури даних для успішного навчання; забезпечуючи стабільний успіх при навчанні з сильно обмеженими наборами даних істини. Розробка таких підходів потребує теоретичного обґрунтування, експериментальної перевірки та оцінки ефективності нових методів та моделей навчання з мінімальною кількістю навчальних даних. Одним із перспективних

напрямів у цій галузі є теорія та методи неконтрольованого навчання, які відповідають усім із цих суттєвих вимог.

Значний внесок у формулювання та розвиток методів неконтрольованого навчання, систем самонавчання та створення інформативних представлень даних внесли роботи Дж. Хінтона, Й. Бенджіо та інших; в галузі аналізу та класифікації даних Інтернет, Ф. Гінголі, Р. Бар-Янай, Т. Гленнана, І. А. Терейковського та багатьох інших; теорії та додатків штучних нейронних мереж, Дж. Хінтона, Г. Цибенко, К. Хорніка, та багатьох інших, в галузі аналізу розподілів даних, включаючи практичні додатки до даних різних типів, Дж. Сандера, М. Естер, К. Фукунагі, П. О. Приставки та багатьох інших.

Розв'язання задачі впевненого розпізнавання класів даних в сценаріях з мінімальними наборами навчальних даних таких як дані трафіку Інтернет дозволить створити більш ефективні інтелектуальні системи з можливими застосуваннями в багатьох областях, включаючи інформаційну безпеку, громадську безпеку, планування та прогнозування, медицину та комерційні програми з зарахуваннями для бюджету.

З огляду на вищезазначене, розробка і дослідження нових ефективних методів забезпечення стабільності та ефективності навчання систем машинного інтелекту з мінімальними даними істини з використанням природної структури даних Інтернет є *актуальною науково-практичною задачею,* що має теоретичне і практичне значення.

Зв'язок роботи з науковими програмами, планами, темами. Результати дисертаційних досліджень реалізовані в рамках держбюджетної теми №247-ДБ19 «Розроблення та виготовлення програмно-апаратних засобів цільового навантаження для повітряного спостереження та альтернативної навігації літального апарату» (№ держреєстрації: 0119U100553), № 421-ДБ22 «Інтелектуалізована система захищеного передавання пакетних даних на базі розвідувально-пошукового безпілотного літального апарату» ((№ держреєстрації: 0122U002361) і виконання тематичних наукових планів досліджень кафедри прикладної математики НАУ 2021 р.

Мета і задачі дослідження. *Метою роботи* є розробка моделей та методів неконтрольованого глибокого навчання процесу розпізнавання даних пакетів трафіку Інтернет, на основі використання структури щільності генеративних представлень для розв'язання задачі впевненого розпізнавання категорій даних Інтернет у додатках, де значні масиви навчальних даних не доступні, або не можуть використовуватись внаслідок значної залежності успіху навчання традиційними методами від джерела даних. Для досягнення поставленої мети необхідно розв’язати такі основні задачі:

1. Провести аналіз сучасних моделей та методів неконтрольованого навчання розпізнавання даних пакетів трафіку Інтернет.

2. Створити теоретичну математичну модель розподілу даних пакетів трафіку Інтернет, включаючи розподіл щільності в просторах генеративних представлень. Провести теоретичне обґрунтування гіпотези про виникнення структурованих представлень моделей генеративного навчання.

3. На основі теоретичної моделі та відомих моделей генеративного навчання розробити та імплементувати моделі неконтрольованого генеративного навчання з даними Інтернет, зокрема, нейромережевих архітектур глибокого навчання.

4. Оцінити, отримати, обробити та підготувати для використання набори даних пакетів трафіку Інтернет та інших типів для навчання генеративних моделей.

5. Провести оцінку та аналіз характеристик розподілу даних пакетів трафіку Інтернет у просторах інформативних генеративних представлень.

6. Розробити, виконати та перевірити методи навчання розпізнавання відомих класів даних пакетів трафіку Інтернет з наборами навчальних даних мінімального обсягу на основі генеративних представлень.

7. Розробити, виконати та перевірити повністю неконтрольовані методи навчання розпізнавання натуральних концептів даних пакетів трафіку Інтернет без вимог даних навчання відомих класів на основі генеративних представлень.

8. Сформулювати, формалізувати та виконати прототипну реалізацію інформаційної технології обробки даних та навчання розпізнавання відомих класів даних пакетів трафіку Інтернет з наборами навчальних даних мінімального обсягу (до кількох зразків) на основі генеративної структури представлень.

Об'єктом дослідження є процеси створення та застосування інформаційних представлень даних Інтернет створених моделями неконтрольованого генеративного навчання.

Предметом дослідження є моделі, методи та засоби створення та аналізу неконтрольованих генеративних представлень даних Інтернет та інших типів, включаючи: нейромережеві моделі неконтрольованого генеративного самонавчання; методи аналізу розподілу даних, включаючи багатовимірні розподіли; методи кластеризації.

Методами дослідження є теорія інформації, статистичні методи, теорія неконтрольованого навчання та представлень – для розробки методів створення інформативних представлень даних Інтернет зі значно зниженою розмірністю; моделі та методи неконтрольованого навчання, включаючи генеративні моделі нейронних мереж автоенкодера – для створення інформативних представлень даних зі значно зниженою розмірністю; методи кластеризації, включаючи кластеризацію за щільністю – для аналізу структури неконтрольованих генеративних представлень; методи статистичного аналізу – для обробки результатів експериментів та верифікації ефективності розроблених методів.

Наукова новизна отриманих результатів:

1. Удосконалено концептуальну і математичну модель розподілів даних пакетів трафіку Інтернет, яка за рахунок використання неконтрольованих генеративних представлень, дозволяє підвищити ефективність виділення інформативних факторів даних.

2. Отримали подальший розвиток методи теорії генеративних представлень у напрямку розробки методів аналізу структури щільності розподілів даних пакетів трафіку Інтернет в представленнях генеративних моделей глибокого навчання, що за рахунок застосування оригінальної архітектури автоенкодера з різким стисненням розмірності шару кодування, забезпечило підвищення ефективності навчання за характеристиками зниження помилки та відтворення даних при неконтрольованому генеративному навчанні з використанням даних Інтернет та інших типів даних.

3. Вперше доведена теорема про категоризацію генеративних представлень, що на підставі методів варіаційного аналізу забезпечує теоретичне обгрунтування методів навчання на основі генеративної структури представлень.

4. Вперше розроблено методи навчання призначені для розпізнавання відомих класів даних пакетів трафіку Інтернет, які за рахунок розроблених концептуальної та математичної моделі, методів теорії генеративних представлень та теореми про категоризацію генеративних представлень, забезпечують: достатню точність розпізнавання; зменшення залежності від джерела отримання навчальних даних; зменшення обсягу навчальних даних в порівнянні з відомими методами контрольованого навчання.

5. *Вперше* розроблено методи навчання призначені для розпізнавання натуральних концептів даних пакетів трафіку Інтернет, які за рахунок визначення структури щільності генеративних представлень, дозволяють реалізовувати розпізнавання

натуральних концептів даних Інтернет без використання маркованих даних, з точністю на рівні відомих методів контрольованого навчання.

6. *Вперше* визначено, формалізовано та виконано прототипну реалізацію інформаційної технології розпізнавання класів даних трафіку Інтернет класу архітектур глибокого навчання, що за рахунок застосування запропонованих методів визначення структури щільності генеративних представлень, дозволяє автоматизувати процес навчання та використання запропонованих моделей і методів розпізнавання класів трафіку Інтернет.

Практичне значення одержаних результатів:

1. Проведено порівняльний аналіз сучасних методів розпізнавання та класифікації даних пакетів трафіку Інтернет. Отримано порівняльні результати ефективності та недоліків розглянутих методів.

2. Отримано програмну реалізацію моделей генеративного навчання автоенкодера з різким стиском розмірності кодуючого шару кодування у програмному середовищі Python, дозволяє отримати інформативні представлення даних Інтернет суттєво зниженої розмірності.

3. Отримано програмне втілення масивів даних пакетів трафіку Інтернет та зображень у програмному середовищі Python.

4. Отримано програмну реалізацію методів визначення структури щільності генеративних представлень у програмному середовищі Python (методи кластеризації за щільністю, багатовимірних гістограм) дозволяє стабільне визначення структури щільності генеративних представлень даних Інтернет, з успішністю генеративного навчання та визначення структури щільності вище 80%.

5. Отримано програмну реалізацію методу навчання розпізнавання відомих класів даних пакетів трафіку Інтернет з використанням структури щільності генеративних представлень у програмному середовищі Python з використанням пакетів та бібліотек машинного навчання. Дозволяє стабільне навчання розпізнавання відомих класів даних Інтернет при зменшення обсягу навчальних даних, у 10–100 разів в порівнянні з відомими методами контрольованого навчання.

6. Отримано програмну реалізацію методу навчання розпізнавання натуральних концептів даних пакетів трафіку Інтернет з використанням структури щільності генеративних представлень, у програмному середовищі Python з використанням пакетів та бібліотек машинного навчання та обробки даних. Дозволяє стабільне навчання розпізнавання натуральних типів (концептів) даних Інтернету без вимоги відомих даних навчання.

7. Отримано програмне прототипне виконання інформаційної технології розпізнавання даних пакетів трафіку Інтернет при наборах навчання мінімального обсягу з використанням структури щільності генеративних представлень у програмному середовищі Python з використанням пакетів та бібліотек машинного навчання.

Результати дисертації використовуються у науково-дослідній діяльності НДЛ протидії кіберзагрозам авіаційної галузі.

Особистий внесок здобувача. Результати що становлять основний зміст дисертації, отримані здобувачем самостійно. У роботах, виконаних у співавторстві, у дисертаційній роботі використовуються результати, що отримані особисто здобувачем: у роботі [2] – здобувачу належить концепція застосування методів та моделей неконтрольованого генеративного самонавчання в системах та процесах обробки даних аеронавігації, участь у написанні тексту, [9] – здобувачу належить концепція, обробка експериментальних даних, участь у написанні та форматуванні тексту, [12] – здобувачу

належить обробка даних, дизайн та виконання моделей машинного навчання, обробка результатів, [14] – здобувачу належить обробка експериментальних даних, підготовка та форматування тексту, [17] – здобувачу належить концепція використання генеративних представлень; участь у написанні та форматуванні тексту.

Апробація результатів дисертації. Результати досліджень дисертаційної роботи доповідалися, обговорювалися та отримали позитивну оцінку на наукових та науково-практичних конференціях: МНТК «Theory and Application of Soft Computing, Computing with Words and Perceptions (ICSCCW)» (Прага, 2019 р.), МНПК «Advances in Computer Science for Engineering and Education (ICCSEEA)» (Київ, 2020 р.), МНТК «Advanced Computer Information Technologies (ACIT)» (Дегендорф, 2020, 2021), МНТК «Information Technologies - Applications and Theory (ITAT)» (Хелпа, 2020, 2021), МНТК Informatics & Data-Driven Medicine (IDDM)» (Ваксйо, 2021), МНПК «Information Society and University Studies (IVUS)» (Каунас, 2021), МНПК «ICT in Education, Research and Industry ICTERI-2021» (Херсон, 2021), МНТК «Soft Computing and Pattern Recognition (SoCPaR)» (MirLabs США, 2022); представлялися на наукових семінарах кафедр прикладної математики та аеронавігації Національного авіаційного університету (2019, 2020 р.); наукових семінарах Київського політехнічного інституту імені І. Сікорського (2019 р.), Ужгородського національного університету (2021 р.).

Публікації. Основні положення дисертації опубліковано у 21 науковій праці, в тому числі в – 1 розділі колективної монографії, 17 наукових статях (16 – у міжнародних рецензованих виданнях, що входять до бази даних Scopus, 1 – у вітчизняних фахових наукових журналах категорії Б), а також 3 матеріалах тез доповідей на конференціях.

Структура та обсяг дисертації. Дисертація складається із анотації, вступу, чотирьох розділів, висновків, додатків, списку використаних джерел і має 129 сторінок основного тексту, 29 рисунків, 20 таблиць. Список використаних джерел містить 110 найменувань. Загальний обсяг роботи складає 139 сторінок.

ОСНОВНА ЧАСТИНА

У вступі обґрунтовано актуальність теми дисертації, сформульовано мету дослідження, визначено коло наукових задач, що належить розв'язати, вказано наукову новизну, практичне значення отриманих результатів, наведено дані про їх апробацію та впровадження.

У першому розділі здійснено огляд та аналіз сучасних методів та моделей неконтрольованого навчання, аналізу розподілів даних у багатовимірних просторах представлень відповідно до мети та гіпотези роботи.

При постановці задач роботи відзначалися проблеми та обмеження при застосуванні традиційних методів і моделей машинного навчання, у випадках та середовищах, де значні набори маркованих навчальних даних або не існують, або не можуть застосовуватися ефективно у зв'язку з сильною залежністю від способів та джерел створення навчальних наборів. Прикладом таких завдань є дані мережі Інтернет, де, як було показано в ряді робіт, успіх навчання традиційними методами може значною мірою залежати від мережі збору даних навчального набору.

З іншого боку, як показують опубліковані результати в галузі систем неконтрольованого навчання, виготовлення неконтрольованих генеративних представлень не вимагає маркованих навчальних даних і представляє логічний та натуральний напрямок розв'язання поставленої задачі. Інформативні структури, які виникають у представленнях генеративних систем навчання можуть використовуватися для підвищення ефективності навчання з мінімальними

навчальними наборами мережевих даних і середовищах із дефіцитом даних навчання або значного зменшення залежності успіху навчання від джерела даних. Діаграма методів та підходів неконтрольованого навчання подано на Рис. 1.

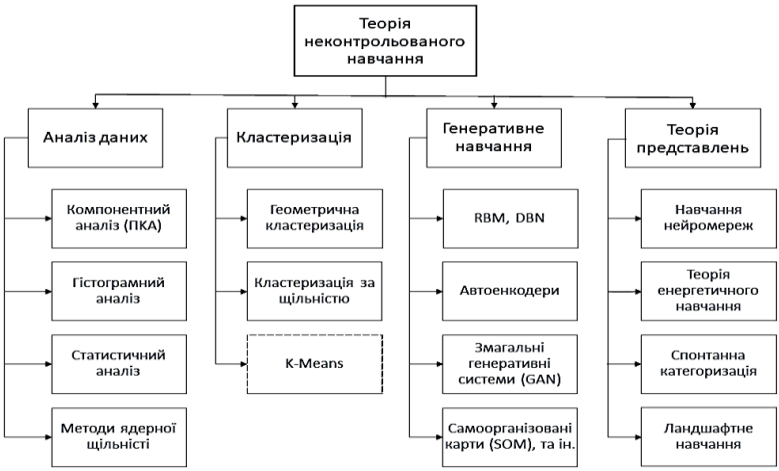

Рис. 1. Методи теорії неконтрольованого навчання

У другому розділі представлено теоретичне обґрунтування методів самонавчання з використанням генеративної структури неконтрольованих представлень (генеративного ландшафту) без значних наборів навчальних даних.

Постановка задачі: припустимо задані набори даних: D, довільного розміру з представленими відомими категоріями $C = \{ C_k \}$ та набір навчальних даних D_0. За умовою завдання або середовища застосування, або a) розмір набору D_0 недостатній для застосування традиційних методів контрольованого навчання; b) навчальні дані набору D_0 не цілком відповідають розподілу вхідних даних (це можливо наприклад у випадках, коли набори D, D_0 отримані з різних джерел). Таким чином, виникає обмеження що застосування стандартних методів контрольованого навчання на наборі D_0 може не забезпечити бажаного рівня точності, як було зазначено у ряді опублікованих результатів у додатках до аналізу даних мереж Інтернет.

Потрібно визначити методи: $M(x, C)$, розпізнавання відомих категорій (класів) даних $C(x) = M(x \in X, C)$ на просторі вхідних даних $D \subset X$, який задовольняє умові прийнятної точності: $e(x \in D) \leq \varepsilon_{max}$, e: середня помилка розпізнавання; та $P(x)$: визначення та розпізнавання натуральних (прихованих) концептів даних, $n(x) = P(x)$, де $N = \{ n_i \}$ – набір натуральних концептів визначений методом на основі набору вхідних даних D, також з умовою прийнятної точності.

Підходи до розв'язання задачі можуть бути встановлені на основі теорії генеративного навчання, яка дозволяє створювати інформативні представлення вхідних даних методами неконтрольованого навчання з мінімізацією помилки відтворення навчальних даних, що не вимагають значних наборів маркованих відомими категоріями.

Припустимо, існує модель генеративного навчання, яка здійснює перетворення кодування та відтворення вихідних даних X у просторі представлення зниженої розмірності R. При виконанні умов 1. Точності відтворення, що вимірюється певною метрикою у вихідному просторі (функція втрат); 2. Зниження розмірності простору представлення, та 3. Загальності, тобто незалежності середньої точності відтворення від конкретного набору даних, значна кількість сучасних опублікованих результатів дають підстави очікувати, що структура розподілу даних у просторі представлення може бути у кореляції з характерними типами вхідних даних представлених вибіркою навчального набору. Далі, значне зниження розмірності простору представлення порівняно з розмірністю вихідних даних значно полегшує розпізнавання інформаційної структури представлення методами кластеризації, у тому числі неконтрольованими, тобто такими що не вимагають зразків відомих концептів. Неконтрольоване навчання генеративних моделей, таких як нейромережеві моделі типу автоенкодера може проводитися стандартними методами навчання, такими як методи зворотного розповсюдження, стохастичного градієнта та іншими, з представницькими наборами даних без асоціації з відомими класами або категоріями.

Теоретичну основу для методів навчання з використанням генеративного неконтрольованого ландшафту представлення закладає *теорема про категоризацію генеративних представлень*, доведена в рамках теоретичного обгрунтування роботи. Її твердження полягає в тому, що за умов генеративного навчання обговорюваних вище, у стані навчальної моделі з мінімальною помилкою відтворення статистично воліють конфігурації генеративних параметрів моделі G і розподілів даних у просторі представлення R з факторизацією областей розподілу характерних категорій даних H_k («прихованих» або «натуральних» концептів):

$$R = E(D) = \sum_{k}^{L} H_k + N;\ \dim O_{nc} = \sum \dim\left(H_k \cap H_j\right) \to min\ ,$$

де $E(D)$ – кодуюче перетворення; H_k – області розподілу прихованих концептів у просторі представлення; N – випадковий шум у даних.

Висновок теореми, доказ якої заснований на варіаційних принципах, вимагає мінімізації розмірності областей перетину прихованих концептів при мінімізації помилки генеративного навчання, і, таким чином, добре визначеної, «категоризованої» структури розподілів концептів в конфігураціях що мінімізують помилку генеративного навчання. Наслідки теореми дозволяють просунутися у розумінні деяких суттєвих питань, наприклад, якою мірою можна стискати інформацію у просторі генеративного представлення, але вона недостатня, щоб перейти безпосередньо до задачі неконтрольованого розпізнавання концептів, оскільки області та параметри розподілу концептів можуть бути невідомі.

Для подальшого просування необхідно зробити додаткові припущення, а саме: простір вхідних даних представлений навчальним набором містить кінцеве число основних концептів зі значним представництвом у наборах, які використовуються для генеративного навчання. Це припущення виправдане у випадках роботи, таких як дані Інтернет або спеціалізовані набори зображень. Це ключове припущення дозволяє очікувати, на підставі теореми категоризації, що простір представлень успішних моделей генеративного самонавчання може мати виражену структуру щільності, з областями підвищеної концентрації даних у районах відповідних латентним регіонам розподілів характерних типів даних (концептів) у просторі представлення:

$$max\, D(y) \sim H_k(y),$$

де $max(D(y))$ – локальні максимуми функції щільності $D(y)$ розподілу даних у просторі представлення; H_k – області розподілу прихованих концептів.

Вищезазначені аргументи дозволяють визначити інформативну структуру представлень – генеративний ландшафт L сформований в результаті успішного генеративного навчання та застосування методів неконтрольованої кластеризації у просторі представлень структурованих, згідно з висновками теореми категоризації. Існують відомі методи кластеризації, такі як кластеризація за щільністю (DbScan, Optics, MeanShift), методи багатовимірних гістограм та інші, що дозволяють аналізувати структури щільності в багатовимірних просторах без відомих зразків концептів, тобто в повністю неконтрольованому режимі.

$$L(D) = \{ d_j \}, j = \overline{1,n} \ = K_d(E(D)),$$

де $\{ d_j \}$ – структури щільності розподілу даних, такі як кластери щільності у просторі генеративного представлення; K_d – алгоритм кластеризації за щільністю.

Методи кластеризації за щільністю для даних D можна визначити як (метод DbScan):

$$n_{\boldsymbol{\varepsilon}}(p) = \{ q \in D \mid dist(p,q) \leq \varepsilon \}$$
$$C_{\boldsymbol{\varepsilon},P_t} = \{ p \in D \mid n_{\varepsilon}(p) \geq P_t \},$$

де ε – радіус околиці (параметр методу); $n_{\boldsymbol{\varepsilon}}(p)$ – функція чисельності околиці; $C_{\boldsymbol{\varepsilon},P_t}$ – ядра щільності.

При досить репрезентативному загальному наборі D, ґрунтуючись на припущенні загальності навчання, можна очікувати, що генеративний ландшафт представлення $L(D)$ описує загальну структуру вхідних даних. Необхідно зазначити, що ні процес генеративного самонавчання моделі, ні визначення генеративної структури представлення не вимагають відомих даних концептів і, таким чином, є повністю неконтрольованими.

На основі визначеної структури щільності (ландшафту) представлень, навіть при мінімальних наборах відомих даних, можна сформувати навчальні набори класів методом вибірки:

$$d_p, d_n(D_c) = \{ d \in L \mid E(D_c) \in d_p \}, \{ d \in L \mid d \neq d\_p \}$$
$$(T_p, T_n) = \left(\{ y \in d_p \}, \{ y \in d_n \} \right),$$

де D_c – набір (зразок) відомих представників класу; d_p, d_n – структури ландшафту, відповідні розподілу класу та даним поза класом; T_p, T_n – набори зразків класу і поза класом у просторі представлення.

Отримані набори використовуються для побудови класифікаторів класів та концептів стандартними методами контрольованого навчання (наприклад, методами опорних векторів, нейронних мереж, найближчого сусіда та ін.). У роботі використовувалися класифікатори типу найближчого сусіда (kNN).

На підставі наведених аргументів, у роботі запропоновані два різновиди методів навчання з використанням генеративного неконтрольованого ландшафту:

1. Повністю неконтрольований метод розпізнавання натуральних концептів даних, що не вимагає відомих даних класів; результатом застосування методу є класифікатори натуральних концептів, що дозволяє відносити вхідні зразки до їх натуральних типів (концептів) $\{K_j\}$. При цьому набір концептів не закладається ззовні, а визначається самим методом.

$$K(x) = K_j\big(E(x)\big),$$

де x – зразок у просторі вхідних даних; $E(x)$ – перетворення кодування; K_j –

класифікатор концепту побудований на підставі генеративної структури представлення.

2. Метод навчання розпізнавання відомих класів з мінімальними наборами навчальних даних на основі генеративного ландшафту представлень. Метод дозволяє класифікувати зразки вихідних даних до одного з відомих класів $\{ C_i \}$ як:

$$C(x) = C_j\big(E(x)\big),$$

де x – зразок у просторі вхідних даних; $E(x)$ – перетворення кодування; C_i – класифікатор відомого класу побудований на підставі генеративної структури представлення та мінімальних навчальних наборів даних D_k в ітераціях навчання.

Методи використовують процеси створення навчальних наборів класів (самплінг) на підставі структури генеративного неконтрольованого ландшафту представлення, як зазначено вище. Детальний опис процесу створення навчальних наборів та побудови класифікаторів концептів представлено у роботі.

Діаграму процесу обробки інформації методами навчання на основі структури щільності генеративних неконтрольованих представлень показано на Рис. 2.

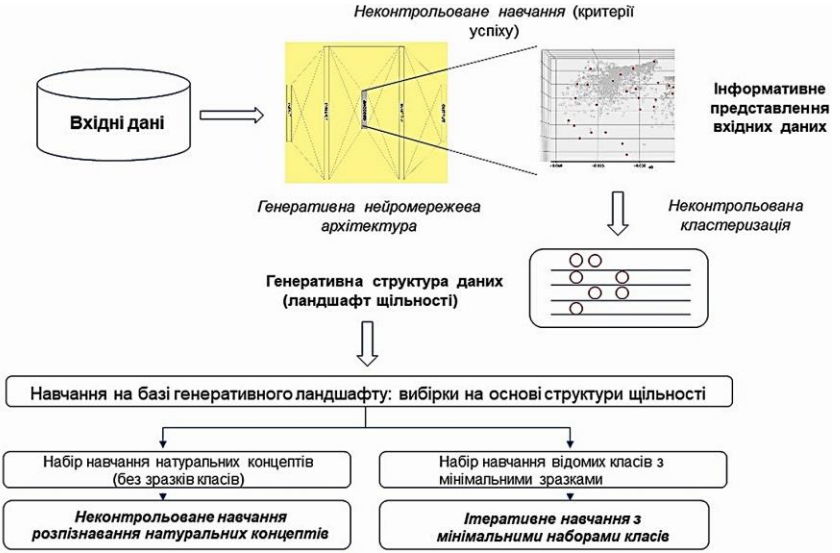

Рис. 2. Блок-схема перетворення інформації при навчанні на основі генеративного ландшафту щільності представлень

Слід зазначити, що генеративна структура представлення описує натуральну структуру розподілу даних і апріорі не можна автоматично укласти що вони будуть повністю відповідати відомим «зовнішнім» класам. Проте, у конкретних випадках включаючи дані Інтернет, які розглядаються в роботі такий висновок може мати підставу, зважаючи на значні відмінності між мережевою поведінкою різних додатків. У цьому випадку, можна припустити, що генеративний ландшафт представлення *L(D)*, визначений повністю неконтрольованими методами може бути використаний для ефективного навчання зі значно зменшеними наборами відомих даних класів даних Інтернет.

Третій розділ містить детальний опис генеративних архітектур, моделей, методів та даних використаних в роботі. На підставі результатів теоретичного обґрунтування, а також опублікованих наукових результатів, була обрана архітектура моделей що використовуються для створення та вивчення генеративних неконтрольованих представлень вхідних даних. Програмна модель, використана в роботі для отримання представлень вхідних даних, являла собою генеративну нейронну мережу типу глибокого симетричного автоенкодера зі значним стиском у центральному шарі представлення, як показано на архітектурній діаграмі Рис. 3.

Вхідні дані моделі являли собою набір векторних даних які визначають сесії Інтернет розміру N та розмірності p, тобто p вхідних параметрів які могли описувати окрему сесію Інтернет (у наборі даних, що використовувався в роботі, p = 22–30) або зображення візуального набору. Кодуючий компонент (Енкодер), який містив шари моделі від вхідного до шару представлення складався з одного шару нейронів розмірності 50–100, з додатковими спеціальними функціями активації (Relu) та нормалізації (Batch Normalization). Генеративний компонент (Генератор) був симетричним Енкодеру.

Інформативне представлення вхідних даних (пакетів трафіку Інтернет та інших типів використаних у роботі) створювалося одним центральним шаром розмірності d = 3–10 (Latent, Рис. 3), тобто координати простору представлення відповідали активаціям латентних нейронів. Таким чином, в результаті обробки вхідного набору ($N \times p$) створювалось маломірне інформативне представлення розмірності d та латентний розподіл даних у просторі представлення розміру ($N \times d$).

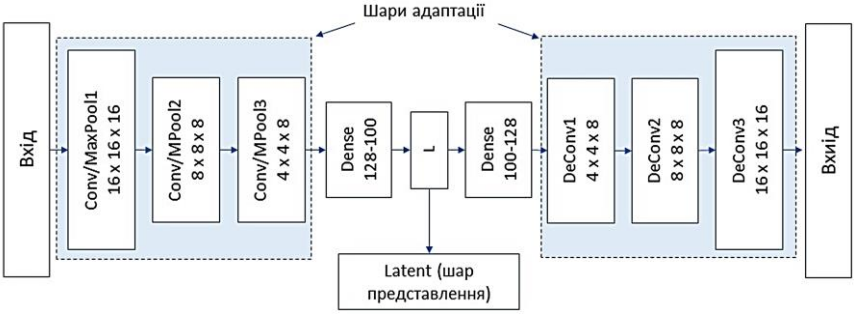

Рис. 3. Архітектурна діаграма моделі глибокого автоенкодера

При роботі з даними зображень, енкодер та генератор моделі додатково містили компоненти фізичної адаптації додані для придбання характеристик зображень різних масштабів (шари адаптації, Рис. 3). Вони складалися зі стандартних у практиці роботи із зображеннями шарів згортки та агрегації по осях координат, 2–3 стадії у кожному з компонентів.

Таким чином, у роботі з даними Інтернет коефіцієнт стиснення даних в результаті створення інформативного представлення у процесі неконтрольованого генеративного навчання становив до 10, з даними зображень, до 10^3. Залежно від розміру прихованих шарів, моделі даних Інтернет мали до 8,000 параметрів, даних зображень до навколо 90,000 параметрів як вказано в Табл. 1. Моделі були реалізовані у програмному середовищі Python з використанням пакетів моделювання нейронних мереж Keras та Tensorflow. Також використовувалися стандартні програмні пакети обробки аналізу та візуалізації даних, такі як numpy, pandas, sklearn-kit, matplotlib та інші.

Таблиця 1

Параметри нейромережевої архітектури моделі глибокого автоенкодера

Шар	Розмірність шару	Інтервал значень	Активація	Формат	Функція помилки
Input	p = 22..30	[0 .. 1]	–	(N, p)	–
Encoder	m = 10..50	[0, ∞[	Relu	(N, m)	–
Latent	d = 3..10	]-∞,+∞[	Leaky Relu	*(N, d)*	–
Output	p	[0 .. 1]	Sigmoid	*(N, p)*	MSE (1)

У процесі створення представлень моделі навчалися стандартними методами навчання нейронних мереж, тобто зворотного розповсюдження, стохастичного зниження за градієнтом та аналогічними. Метою генеративного навчання було зниження помилки відтворення, тобто метрики відхилення вхідного набору від генеративного відтворення створеного моделлю в ітераціях неконтрольованого навчання.

Дані для навчання генеративних моделей та створення генеративних представлень: Відповідно до задачі роботи, дані представляли інтернет-трафік, отриманий із записів пакетів трафіку протоколу Інтернет (packet dumps) телекомунікаційної мережі загального призначення доступних на публічному сервері WITS університету Вайкато, Нова Зеландія.

Інтернет-пакети з отриманих записів мережевого трафіку були зібрані в сесії з як детально визначено у роботі. У результаті було отримано набір Інтернет-сесій розміру близько 130 тисяч сесій який представляв до 1,000 характеристичних типів сесій та додатків. Інтернет сесії були елементами немаркованого набору даних генеративного навчання який використовувався для навчання нейромережевих моделей та створення інформативних представлень даних і визначався числовим вектором 22–30 параметрів темпоральної та об'ємної статистики пакетів даних у сесії, наприклад розмір пакетів, інтервали між пакетами в сесіях як зазначено в Табл. 2.

Таблиця 2

Параметри набору даних пакетів трафіку Інтернет

Тип параметрів	Число параметрів	Опис	Тип даних	Попередня обробка
Загальні	6	Загальна тривалість; загальний розмір (по напрямку), кількість пакетів (по напрямку), протокол	Ціле число, R	Масштабування, інтервал [0,1]
Статистика розміру пакету	8–12	Мін, макс, середнє, стандартне відхилення, ентропія розподілу пакетів за розміром	R	Масштабування
Статистика часового інтервалу	8–12	Мін, макс, середнє, стандартне відхилення, ентропія розподілу пакетів за часовим інтервалом	R	Масштабування

Частина масиву було марковано класами додатків найбільш представлених у наборі на основі відомого порту, який є параметром інтернет-протоколу. Треба відзначити що марковані зразки використовувалися тільки для вивчення структури представлень та при розробці та аналізі методів класифікації, в той час, як навчання генеративних моделей та визначення генеративної структури представлень проводилося в повністю неконтрольованому процесі без маркованих даних.

Для перевірки загальності результатів та методів, у роботі також використо-вувався масив даних зображень місцевості, отриманий методами аерофотогра-фування. Масив

який містив близько 10,000 зображень, розподілених за класами типів місцевості такі як ліс, поле, водоймище та інші) детально описаний у роботі та у кількох опублікованих наукових працях.

У розділі також визначено та детально описано методи аналізу розподілів даних в інформаційних представленнях генеративних моделей.

У **четвертому** розділі наведено результати експериментальної частини роботи. Їх можна поділити на три логічні групи:

1. Вимірювання та аналіз розподілів даних та структури генеративних представлень;
2. Експериментальна перевірка методів неконтрольованого визначення структури щільності генеративних представлень (генеративного ландшафту представлень);
3. Експериментальна перевірка методів навчання на генеративному ландшафті розроблених у роботі.

За результатами аналізу розподілів даних у генеративних представленнях спостерігалася виражена структура представлень корельована з додатками Інтернет (Рис. 4) що також підтверджується спостереженнями виникнення вираженої генеративної структури у процесі неконтрольованого генеративного навчання нижче.

Отримані результати дають експериментальне підтвердження висновків теоретичної частини роботи про виникнення про виникнення концептуальних структур представлень у процесі генеративного неконтрольованого навчання.

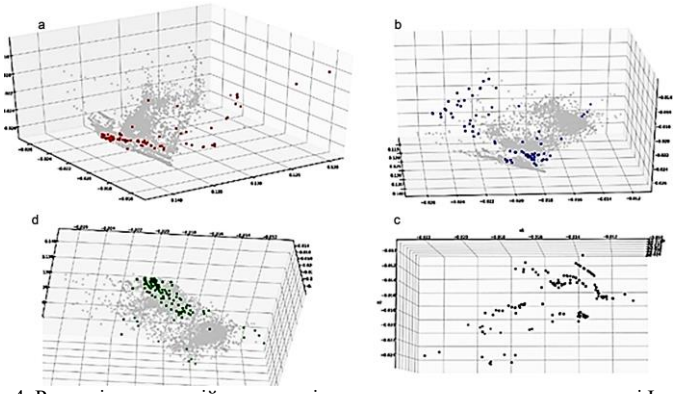

Рис. 4. Розподіли категорій у просторі генеративного представлення, дані Інтернет (за годинниковою стрілкою, a - d, Інтернет програми: Месенджер, Торрент, протокол HTTP, Онлайн гра, на тлі загального розподілу, сірий колір)

Ці спостереження були підтверджені результатами експериментів з аналізу структури латентних представлень що виникають в процесі генеративного навчання і класифікації в просторі генеративних представлень (Табл. 3).

Таблиця 3

Структуризація у просторі представлень при генеративном навчанні

Епоха навчання	**Кількість структур (кластерів) щільності**	**Точність навчання [(1)]**
0 (ненавчена модель)	5	0.75
20	8	0.908
40	10	0.910
60	12	0.914
80	12	0.913

[(1)] Метрика точності F1, класифікатор типу kNN.

В експериментах з навчанням генеративних моделей були відзначені: по-перше, виникнення помітної структури щільності в генеративних представленнях моделей у процесі генеративного навчання; по-друге, стійке підвищення точності класифікації при навчанні на зразках у просторі генеративного представлення у процесі генеративного навчання моделей підтверджене вимірюванням параметрів розподілів концептів у просторах представлення та точності класифікації; і також, рівень успіху навчання генеративних моделей, вище 80%.

Отримані результати підтверджують ключове положення роботи про наявність кореляційного зв'язку між структурою що виникає у генеративних представленнях у процесі успішного неконтрольованого генеративного навчання та характеристиками розподілу головних категорій вхідних даних. В результаті було експериментально підтверджено висновки теореми про категоризацію генеративних представлень у теоретичній частині роботи та практично обґрунтовано рішення щодо вибору архітектури та параметрів генеративних моделей навчання. Аналогічні результати були отримані з набором даних зображень що підтверджує загальний характер спостережуваного ефекту категоризації при генеративному навчанні у повній відповідності до висновків теоретичної частини роботи.

Методи навчання з використанням генеративної структури (ландшафту) представлень

У роботі проведено детальну експериментальну перевірку методів навчання на генеративному ландшафті з наборами даних Інтернет та зображень. Результати експериментальної перевірки методів за точністю навчання та розпізнавання представлені у Табл. 4.

Таблиця 4

Експериментальна перевірка методів навчання

Клас	**Початкова точність**	**Кінцева точність** [(1)]
Розпізнавання відомих класів з мінімальними навчальними наборами (набір Інтернет)		
Месенджер	0.748	0.921
E-мейл	0.702	0.930
Відео / аудіо стрім	0.798	0.892
DNS	0.861	0.915
Розпізнавання натуральних категорій (набір зображень)		
Натуральна категорія	**Щільність області розподілу** [(2)]	**Кінцева точність** [(1)]
“Ліс-вода” (0)	1.5×10^3	0.980
“Поле” (1)	1.0×10^3	0.977
“Вода” (2)	1.1×10^3	0.984
“Структури” (6)	380	0.949
“Дороги” (16)	570	0.956

[(1)] Точність класифікатора типу kNN (найближчого сусіда) на наборі 100 маркованих образців після 10 ітерацій навчання. Метрика точності F1.
[(2)] Відносно середньої щільності розподілу загального набору, 1.

Наведені результати підтверджують ефективність розроблених методів для розв'язання поставленої задачі роботи. Результати за точністю навчання з мінімальними наборами даних з набором даних зображень наведені в роботі підтверджують загальний характер запропонованих методів та можливість їх використання з різними типами даних у різних галузях використання.

Порівняльний аналіз

У роботі проведено порівняльний аналіз застосування методу навчання на основі генеративної структури представлень з результатами навчання класифікації даних Інтернет опублікованими в літературі (Табл. 5).

Таблиця 5

Порівняльні результати з методами класифікації трафіку Інтернет

Клас Інтернет	Ландшафтний метод, початкова / кінцева точність [1]	Методи контрольованого навчання [2]
Месенджер, мейл	0.75 / 0.92	0.6 – 0.84
Інтернет-протоколи	0.86 / 0.92	0.78 – 0.91
Загально	0.78 / 0.91	0.75 – 0.92

[1] Розмір навчального набору, початкова точність: 3–10 позитивних зразків класу; кінцева точність: 100 зразків класу (10 ітерацій навчання); метрика точності F1.
[2] Набір навчання відрізнявся від набору верифікації; набір навчання від сотень до кількох тисяч зразків класів.

З результатів порівняння точності класифікації у Таблиці 5 можна зробити висновок що методи навчання з використанням генеративного ландшафту дозволяють значно зменшити залежність результатів навчання від вибору навчального набору та досягати результатів навчання порівнянних із результатами традиційних методів контрольованого навчання зі значно меншими навчальними наборами.

Результати експериментальної частини роботи, у тому числі експериментальна перевірка методів навчання на ландшафті дозволяють визначити *інформаційну технологію навчання з мінімальними даними навчання з використанням генеративної структури (ландшафту) представлень* зведенням та формалізацією процесів навчання з використанням генеративного ландшафту описаних у теоретичній та експериментальній частинах роботи. Технологія поєднує в єдиний загальний процес етапи:

– підготовки та попередньої обробки даних;

– енеративного навчання та створення генеративних представлень;

– визначення структури генеративних представлень методами неконтрольованої кластеризації;

– створення навчальних наборів класів та концептів на основі виявленої генеративної структури представлення;

– навчання класифікаторів класів на основі генеративного ландшафту в ітераціях навчання.

Таким чином, технологія узагальнює процеси навчання на основі генеративної структури представлень для даних різних типів та сфер застосування. Технологія вимагає мінімальних наборів відомих даних навчання, до окремих зразків відомих класів (few shot learning).

При перевірці з класами даних Інтернет та зображень застосування технології дозволило досягти результатів за точністю розпізнавання практично порівняних з опублікованими результатами сучасних методів при використанні даних загальних мереж (тобто які можуть відрізнятися від мережі-джерела даних навчання), з мінімальними наборами даних (при 30–100 відомих зразків, проти тисяч при використанні традиційних методів контрольованого навчання).

При цьому результат навчання не залежить від якості навчальних наборів отриманих з інших мереж. Навіть з одиничними відомими зразками класів, технологія навчання з використанням генеративного ландшафту представлень дозволяє досягти точності розпізнавання класів значно вище випадкової. Методи використані в запропонованій технології були перевірені з наборами даних різних типів та галузь застосування: дані Інтернет та зображень місцевості.

Діаграму інформаційних процесів розробленої інформаційної технології навчання з використанням генеративного ландшафту представлень наведено у рис. 5. Діаграма

містить детальний опис етапів, процесів та результатів обробки даних при застосуванні технології.

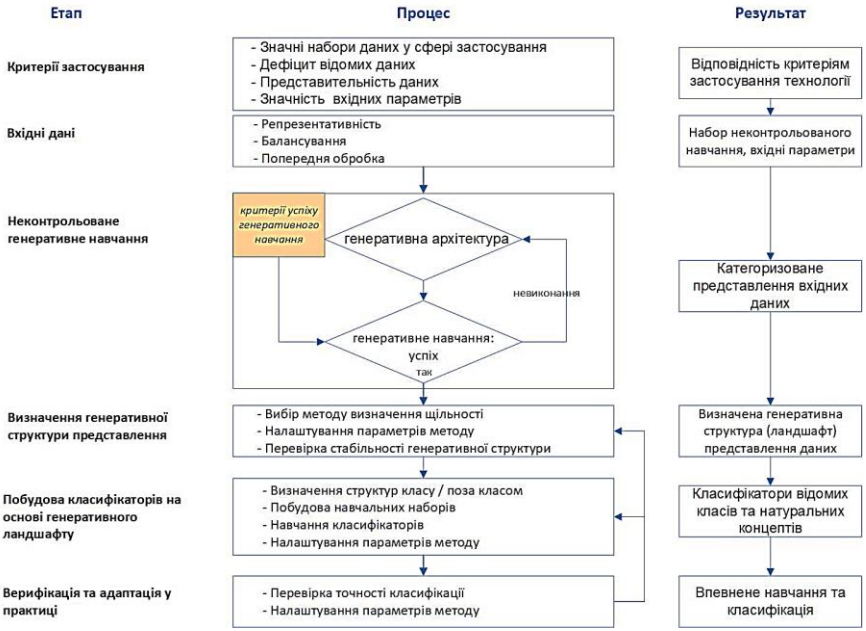

Рис. 5. Інформаційна технологія навчання з використанням генеративного ландшафту представлень

Застосування запропонованої технології дозволяє досягти високого рівня точності розпізнавання натуральних категорій без відомих даних і може застосовуватися та мати важливе значення при аналізі даних невідомої структури та/або походження.

Таким чином, інформаційна технологія навчання з використанням генеративної структури даних розроблена та перевірена в роботі дозволяє узагальнити визначення методів навчання з використанням неконтрольованого генеративного ландшафту представлень даних Інтернет запропонованих та перевірених у роботі на інші типи даних та галузі застосування стандартним повністю визначеним процесом. Результати теоретичних та експериментальних досліджень роботи підтверджують потенціал використання технології у широкому діапазоні задач та галузь з різними типами вхідних даних.

ВИСНОВКИ

1. Удосконалено математичну модель розподілів даних пакетів трафіку Інтернет у генеративних представленнях завдяки теоретичному аналізу зв'язку між геометричними властивостями розподілів і ефективністю навчання генеративних моделей, що дозволило забезпечити теоретичну основу методів навчання на основі генеративного ландшафту представлень та підвищити ефективність виділення інформативних факторів даних.

2. Розроблено та реалізовано оригінальні нейромережеві моделі глибокого неконтрольованого генеративного навчання типу автоенкодера з різким стиском

латентного шару в середовищі програмування Python що дозволило створювати інформативні представлення даних пакетів трафіку Інтернет низької розмірності зі збереженням ключових характеристик розподілу даних.

3. Доведена теорема про категоризацію генеративних представлень (при певних припущеннях та умовах), що на підставі методів варіаційного аналізу забезпечує теоретичне обґрунтування методів навчання на основі генеративної структури представлень даних Інтернет та інших типів.

4. Отримано, оброблено та підготовлено до використання навчальні набори даних Інтернет та візуальних даних у середовищі програмування Python що дозволило стабільно вивчати моделі та аналізувати генеративні представлення даних пакетів трафіку Інтернет.

5. Знайдено розв'язання задачі визначення структури щільності генеративних представлень неконтрольованими методами без вимоги відомих даних навчання засноване на теоретичних основах процесу створення структурованих генеративних представлень та застосування методів неконтрольованої кластеризації, багатовимірних гістограм що дозволило підвищити стабільність розпізнавання структур щільності генеративних представлень від рівня теоретичної можливості до рівня стабільного використання в інформаційній технології з успішністю генеративного навчання та визначення структури щільності вище 80%.

6. Розв'язано задачу стабільного навчання розпізнавання відомих класів даних пакетів трафіку Інтернет на основі структури щільності генеративних представлень, що за рахунок розроблених концептуальної та математичної моделі, методів теорії генеративних представлень та теореми про категоризацію генеративних представлень забезпечують: достатню точність розпізнавання, на рівні відомих методів контрольованого навчання; зменшення залежності від джерела отримання навчальних даних; зменшення обсягу навчальних даних, у 10–100 разів в порівнянні з відомими методами контрольованого навчання. Отримано програмну реалізацію методів в середовищі програмування Python на основі сучасних пакетів та бібліотек глибокого навчання, машинного навчання, аналізу та обробки даних.

7. Розроблено повністю неконтрольовані методи розпізнавання натуральних концептів даних пакетів трафіку Інтернет, без вимог даних навчання відомих класів на основі структури щільності генеративних представлень, з точністю на рівні відомих методів контрольованого навчання. Отримано програмну реалізацію методів в середовищі програмування Python на основі сучасних пакетів та бібліотек глибокого навчання, машинного навчання, аналізу та обробки даних.

8. Визначено, формалізовано та виконано концептну реалізацію інформаційної технології розпізнавання класів трафіку Інтернет класу архітектур глибокого навчання на основі запропонованих методів визначення структури щільності генеративних представлень, що дозволяє автоматизувати процес навчання та використання запропонованих моделей і методів розпізнавання класів трафіку Інтернет з мінімальними навчальними наборами при збереженні або підвішенні точності розпізнавання під час використання з данними загальних мереж Інтернет в порівнянні з методами контрольованого навчання. Отримано програмне концептне виконання інформаційної технології навчання розпізнаванню даних пакетів трафіку Інтернет у програмному середовищі Python з використанням пакетів та бібліотек машинного навчання.

Запропоновані в роботі методи мають низку переваг порівняно з сучасними підходами у галузі розпізнавання класів даних Інтернет: точності розпізнавання відомих класів, порівнянна або перевершуюча результати відомих методів

контрольованого навчання при класифікації даних загального джерела; вимоги до навчальних наборів, необхідні розміри наборів зменшено в 10–100 разів; стабільності навчання, успіх навчання не залежить від джерела навчального набору; універсальності застосування, ефективність методів підтверджена з наборами даних суттєво різних за характером та джерелом.

За результатами роботи розв'язано поставлену задачу впевненого розпізнавання даних пакетів трафіку Інтернет та інших типів зі зменшеними наборами навчальних даних відомих класів та суттєвого зменшення залежності успіху навчання від джерела даних. Розроблені у роботі методи та інформаційна технологія навчання на основі генеративної структури інформативних представлень показали значну ефективність під час навчання з мінімальними наборами навчальних даних. При цьому процес навчання є гнучким та ітеративним з можливістю навчання у міру емпіричного досвіду та/або наявності навчальних даних. Успішне застосування методів неконтрольованого навчання для розв'язання задачі роботи, а також експериментальні результати роботи з перевірки розроблених методів з даними суттєво різних типів дають підставу розраховувати що методи та інформаційна технологія запропоновані та перевірені в роботі можуть бути використані в ширшому колі завдань та додатків зі різними типами даних і задач розпізнавання.

ПУБЛІКАЦІЇ ЗДОБУВАЧА ЗА ТЕМОЮ ДИСЕРТАЦІЇ

1. Dolgikh S., «Spontaneous concept learning with deep autoencoder». *International Journal of Computational Intelligence Systems*, 12 (1), pp. 1–12, 2018. (*Scopus, Q1*)
2. Shmelyova T., Sterenharz A., Dolgikh S., «Artificial Intelligence in Aviation Industries: methodologies, education, applications and opportunities». IGI Global, 2019. *(розділ в колективній монографії)*
3. Dolgikh S. «Spontaneous categorization and self-learning with deep autoencoder models», *Advances in Aerospace Technology (Proceedings of National Aviation University)*, 2019, 80 (3), pp. 51-60. https://doi.org/10.18372/2306-1472.80.14274 *(категорія Б)*
4. Dolgikh S., «Low-dimensional representations in generative self-learning models». *CEUR Workshop Proceedings*, 2020, vol. 2718, pp. 239-245. *(Scopus)*
5. Dolgikh S., «Identifying explosive epidemiological cases with unsupervised machine learning». *CEUR Workshop Proceedings*, 2020, vol. 2753, pp. 1-10. *(Scopus)*
6. Dolgikh S., «Topology of conceptual representations in unsupervised generative models». *CEUR Workshop Proceedings*, 2021, vol. 2915, pp. 150-157. *(Scopus)*
7. Dolgikh S., «Sparsity constraint in unsupervised concept learning». *CEUR Workshop Proceedings*, 2021, vol. 2962, pp. 188-194. *(Scopus)*
8. Dolgikh S. «Generative conceptual representations and semantic communications». *International Journal of Computer Information Systems and Industrial Management Applications*, 14, 2022, pp. 239-248. *(Scopus)*
9. Prystavka P., Dolgikh S., Cholyshkina O., Kozachuk O., «Latent representations of terrain in aerial image classification», *CEUR Workshop Proceedings*, 2021, vol. 3013, pp. 86-95. *(Scopus)*
10. Dolgikh S., «Unsupervised clustering in epidemiological factor analysis». *The Open Bioinformatics Journal*, 14 (1), 2021, pp. 63-72. DOI: 10.2174/1875036202114010063. (*Scopus*)
11. Dolgikh, S. «Categorized representations and general learning». 10th International Conference on Theory and Application of Soft Computing, Computing with Words and Perceptions ICSCCW 2019. Advances in Intelligent Systems and Computing, 2020, vol.

1095. Springer, Cham. (Online ISBN 978-3-030-35249-3) pp. 93-100 https://doi.org/10.1007/978-3-030-35249-3_11 *(Scopus)*

12. Seddigh N., Nandy B., Bennett D., Ren Y., Dolgikh S. et al., «A framework & system for classification of encrypted network traffic using Machine Learning». 15th International Conference on Network and Service Management (CNSM), 2019, p. 1-5, doi: 10.23919/CNSM46954.2019.9012662. (Electronic ISSN: 2165-963X) *(Scopus)*
13. Dolgikh S., «On unsupervised categorization in deep autoencoder models». Advances in Computer Science for Engineering and Education III. ICCSEEA-2020. Advances in Intelligent Systems and Computing, 2021, vol. 1247, Springer, Cham. (Online ISBN 978-3-030-55506-1), p.155-166. https://doi.org/10.1007/978-3-030-55506-1_23 *(Scopus)*
14. Prystavka P., Cholyshkina O., Dolgikh S., Karpenko D., «Automated object recognition system based on convolutional autoencoder». 10th International Conference on Advanced Computer Information Technologies (ACIT), 2020, p. 830-833. (Electronic ISBN:978-1-7281-6760-2) doi: 10.1109/ACIT49673.2020.9208945. *(Scopus)*
15. Dolgikh S., «Native concept frameworks in unsupervised generative learning». 11th International Conference on Advanced Computer Information Technologies (ACIT), 2021, pp. 748-752, (Electronic ISBN:978-1-6654-1854-6) doi: 10.1109/ACIT52158.2021.9548372. *(Scopus)*
16. Dolgikh S. «Synchronized conceptual representations in unsupervised generative learning». Proceedings of the 13th International Conference on Soft Computing and Pattern Recognition (SoCPaR 2021), Lecture Notes in Networks and Systems, vol. 417 Springer (Online ISBN 978-3-030-96302-6), 2022, p. 23-32. https://doi.org/10.1007/978-3-030-96302-6_2 (*Scopus*)
17. Prystavka P., Dolgikh S., Kozachuk O., «Terrain image recognition with unsupervised generative representations: the effect of anomalies», 12th International Conference on Advanced Computer Information Technologies (ACIT, Ruzomberok, Slovakia), 2022, p. 485-488. *(Scopus)*
18. Doligkh S., «Categorization in unsupervised generative self-learning systems». *International Journal of Modern Education and Computer Science*, 13 (3), 2021, p. 68-78. DOI: 10.5815/ijmecs.2021.03.06 (*Scopus*)
19. Dolgikh S., «Unsupervised landscape, complex observations and association learning». 5th International Conference «Computational Intelligence» IntSol-2019, 2019, p.145-147.
20. Dolgikh S., «Parameter-less histogram-scale method of bandwidth estimation in density based clustering». Матеріали XV міжнародної конференції «Контроль і управління в складних системах (КУСС-2020)», м. Вінниця, 8-10 жовтня 2020 р. – Електрон. текст. дані. – Вінниця: ВНТУ, 2020. – Режим доступу: http://ir.lib.vntu.edu.ua//handle/123456789/30662.
21. Dolgikh S., «Characteristics of categorized latent representations in unsupervised generative Learning». 9th World Congress «Aviation in the XXI Century» National Aviation University (Київ, Україна) 2020. https://conference.nau.edu.ua/index.php/Congress/Congress2020/ paper/viewFile/7602/6495

АНОТАЦІЯ

Долгих С. М. Інформаційна технологія розпізнавання мережевих даних Інтернет на основі генеративних нейромережевих моделей – Рукопис.

Дисертація на здобуття наукового ступеня кандидата технічних наук за спеціальністю 05.13.06 – «Інформаційні технології». – Національний авіаційний університет, Київ, 2023.

У роботі проведено дослідження, теоретичне обґрунтування, розробку, програмне виконання та експериментальну перевірку методів навчання розпізнавання класів даних пакетів трафіку Інтернет та інших типів з даними навчання мінімального обсягу на основі структури щільності генеративних представлень даних та запропоновано інформаційну технологію обробки даних та навчання машинних систем на основі структури щільності генеративних представлень.

У теоретичній частині роботи досліджувалися методи створення інформативних генеративних представлень та доведено теорему про категоризацію в генеративних представленнях, що лежить в основі методів навчання з мінімальними наборами відомих даних, запропонованих у роботі.

На підставі результатів теоретичної частини та огляду сучасних методів та моделей навчання штучних систем, запропоновані методи навчання з використанням неконтрольованої генеративної структури (ландшафту щільності) представлень даних Інтернет: метод виявлення характерних типів даних без вимог відомих даних; та метод ітеративного навчання на генеративному ландшафті з мінімальними наборами навчальних даних, до кількох зразків. На основі результатів теоретичних досліджень та експериментальної перевірки запропонованих методів запропоновано інформаційну технологію навчання з використанням неконтрольованої генеративної структури (ландшафту щільності) представлень, яка з'єднує обробку даних, навчання генеративних моделей та виявлення інформаційної структури даних у єдиний процес, який може застосовуватися з даними різних джерел та типів.

Результати роботи підтверджуються ретельним аналізом теоретичних основ, досконалою експериментальною перевіркою та рецензованими публікаціями в українських та міжнародних наукових виданнях.

Ключові слова: інформаційна технологія, теорія неконтрольованого навчання, теорія генеративних представлень, кластеризація, штучні нейронні мережі, моделі глибокого навчання.

ABSTRACT

Dolgikh S. Information technology of learning and classification of Internet packet traffic data based on generative neural models. – Manuscript**.**

Dissertation for the degree of Candidate of Technical Sciences degree in specialization 05.13.06 "Information technologies" – National Aviation University, Kyiv, 2023.

In the thesis, a research into theoretical foundations of unsupervised generative learning, architecture of generative models, design and development, implementation and experimental verification was carried out to propose and verify methods and an information technology of training machine intelligence systems with minimal sets of known data based on generative density structure (landscape) of informative representations created by generative models in the process of unsupervised training with minimization of generative error. Developing such methods is an essential challenge in a number of critical applications including analysis and classification of data in computer networks and Internet. As was established in a number of studies, applying conventional methods with standard sets of training data can affect generality and accuracy of methods in practical applications where data in the networks differs significantly from the sources of training data.

The proposed methods are based on the informative structure of unsupervised generative representations produced with models of generative self-learning that do not require known data to produce. Completely unsupervised methods of determination of generative structure of informative representations proposed and verified in the thesis can produce additional

essential information about the input distributions to a learning model and allow to significantly reduce the requirement for known data to achieve confident learning of both externally known classes and the common general types or “natural concepts” in the data, offering a natural solution to the identified challenges in the stated problem of Internet traffic classification.

In the theoretical part of the thesis, methods of creating informative generative representations were investigated and a theorem of categorization in generative representations proven under a number of identified conditions. The theorem establishes a theoretical foundation for introduction and definition of methods of learning characteristic types (native concepts) and known classes of Internet packet data with minimal sets of training samples based on the density cluster structure in the latent distributions of data proposed and developed in the thesis. The methods use the cluster structure of density distributions in the informative low-dimensional generative representations of Internet packet data, created in the process of unsupervised generative learning to produce latent samples associated with natural concepts or a known classes of interest and construct classifiers of classes and natural concepts with improved accuracy results and reduced dependency on the significant amounts of training data.

The proposed approach has a number of essential advantages compared to conventional supervised methods of machine intelligence, including: flexibility, in learning specific classes and concepts of interest without the constraints of confident knowledge of the complete conceptual structure of the data; the ability to learn iteratively, starting with minimal known samples (down to a handful of samples) and improve learning results when new data becomes available without full retraining of the generative model; massively reduced requirement for prior known training data; and, in a strong correspondence to the stated problem of the thesis, reduce to the minimum the dependence of the learning success on the source of training data via employing natural generative structure of the latent distributions of the data in the network. As well, the proposed methods have interesting parallels to learning of biological systems that is characterized by flexibility and ability to learn successfully with minimal data as and when it becomes available.

On the base of methods proposed and verified in the thesis, the information technology of minimal sample learning based on density structure (landscape) of informative generative representations was developed. The technology combines the stages of: data processing; selection and training of generative models in an unsupervised process; determination of the density structure of latent representations and learning based on the identified generative structure (landscape) of generative representations into a single information process that can be generalized and extended to data of different types and origin in different domains and problem areas.

The results of the thesis are supported by a thorough review of the theoretical foundations of the problem and the existing approaches in Internet data analysis and classification, comprehensive design of the models based on solid theoretical foundations, extensive and comprehensive experimental verification; presentation and positive acceptance of the results by the research community at international and Ukrainian scientific conferences and seminars; and peer-reviewed publications in Ukrainian and international scientific literature.

Keywords: information technology; unsupervised learning theory, theory of generative representations; clustering; artificial neural networks, deep learning.