

ЗАСТОСУВАННЯ ГЕНЕРАТОРІВ ПСЕВДОВИПАДКОВИХ ЧИСЕЛ У МАШИННОМУ НАВЧАННІ

О. І. Гарасимчук¹, С. С. Войтусік², О. А. Цебак²

Національний університет “Львівська політехніка”,

¹ кафедра захисту інформації,

² кафедра безпеки інформаційних технологій

E-mail: oleh.i.harasymchuk@lpnu.ua, stepan.s.voitusik@lpnu.ua, oleh.a.tsebak@lpnu.ua

© Гарасимчук О. І., Войтусік С. С., Цебак О. А., 2025

Проаналізовано роль генераторів псевдовипадкових чисел у забезпеченні ефективності та стабільності алгоритмів машинного навчання. Визначені ключові сфери і форми такого застосування.

Особливу увагу приділено впливу якості генераторів псевдовипадкових чисел (ГПВЧ) на продуктивність моделей, зокрема в задачах ініціалізації нейронних мереж, синтетичного збільшення даних, стохастичної оптимізації та відтворюваності результатів. Розглянуто проблеми періодичності, передбачуваності та низької адаптивності генераторів до специфічних задач ML. Проаналізовано сучасні підходи до оцінки якості ГПВЧ, зокрема емпіричні, теоретичні та прикладні тести. Дослідження також висвітлює перспективи використання глибокого навчання для аналізу ГПВЧ і виявлення закономірностей у псевдовипадкових послідовностях.

Запропоновано рекомендації щодо підвищення ефективності систем машинного навчання шляхом оптимізації використання ГПВЧ. Результати дослідження сприяють глибшому розумінню ролі ГПВЧ у сучасному машинному навчанні та визначають нові напрями для покращення продуктивності моделей.

Ключові слова: генератори псевдовипадкових чисел, машинне навчання, ініціалізація нейронних мереж, стохастична оптимізація, відтворюваність.

Вступ

У сучасному цифровому світі, що насичений інформацією та науковими інноваціями, генератори псевдовипадкових чисел та генератори псевдовипадкових послідовностей (ГПВП) мають широке використання у різноманітних сферах діяльності. Вони застосовуються для моделювання процесів у науці, економіці, військовій справі та інших галузях, а також для створення ключів у криптографічних системах, що забезпечують захист фінансових транзакцій і конфіденційність даних у банківських установах [1–4]. Ці генератори поєднують властивості випадковості та детермінованості, що робить їх важливими інструментами в багатьох критичних сферах.

Машинне навчання (ML) посідає важливе місце в розвитку технологій, забезпечуючи вдосконалення пристроїв та послуг через алгоритми, які навчаються на даних. Від штучного інтелекту до автономних систем, ML стимулює інновації в різних галузях. Сучасні тренди вказують на розвиток більш автономних та ефективних технологій, зокрема через автоматизацію процесів у

межах AutoML, що дає змогу створювати надійні моделі без глибоких технічних знань. Нові напрями, як *tinyML*, відкривають можливості для інтеграції машинного навчання в мікроконтролери та пристрої Інтернету речей, що покращує прийняття рішень та оптимізацію процесів.

З розвитком машинного навчання та його застосуванням у різних сферах, зокрема в медицині, економіці та технологіях, зростають вимоги до точності та стабільності моделей, що безпосередньо залежать від якості ГПВЧ. Низькоякісні генератори можуть призводити до проблем з повторюваністю експериментів та нестабільністю результатів, що ускладнює наукові дослідження та практичне впровадження моделей. Псевдовипадкові числа відіграють критичну роль у навчанні моделей, зокрема під час ініціалізації нейронних мереж, де вони сприяють уникненню перенавчання і покращують адаптацію до даних. Якість генерованих чисел впливає на ефективність процесу навчання та точність результатів, що робить їх важливим фактором у забезпеченні високої продуктивності алгоритмів машинного навчання. Саме тому ця праця є надзвичайно актуальною на сьогодні.

Огляд літературних джерел

Генератори псевдовипадкових чисел і послідовностей, окрім моделювання процесів, відіграють важливу роль у забезпеченні безпеки інформації, адже з їх допомогою розробляються надійні криптографічні методи захисту [5–8]. Без цих технологій було б значно складніше гарантувати цілісність та конфіденційність інформації в умовах постійної кіберзагрози. Сучасні розробки в цій галузі значно підвищують ефективність систем безпеки, роблячи їх критично важливими для захисту даних у багатьох сферах.

Генератори псевдовипадкових чисел – це алгоритми, які створюють числа, що імітують випадковість, хоча насправді вони є детермінованими [9–11].

Якість генератора прямо залежить від складності його алгоритму та надійності застосовуваних моделей. Якщо алгоритм недостатньо складний або передбачуваний, зломисники можуть використовувати послідовність чисел для відновлення алгоритму. Тому одним із важливих аспектів є забезпечення того, щоб алгоритм ГПВЧ був стійким до інверсії, що значно ускладнює злам. Однак навіть найбільш складні алгоритми можуть бути виявлені або відтворені, якщо зібрати достатньо довгу послідовність згенерованих чисел, що створює потенційні ризики для використання таких систем у критичних застосунках.

Продуктивність таких генераторів, потребує довготривалої роботи для досягнення високої якості, що можна перевірити за допомогою тестів на випадковість, таких як батарея тестів Die Hard, доступних через пакет програм Dieharder. Крім того, важливо, щоб PRNG був складним для інверсії, оскільки це може дозволити з'ясувати майбутні значення на основі попередніх, що знижує рівень безпеки.

Тестування генераторів псевдовипадкових чисел може базуватися *на таких типах перевірок* [2, 3, 12–16]:

- *емпіричні*, які зазвичай передбачають аналіз послідовностей чисел, згенерованих машинними алгоритмами, і оцінку їх якості через встановлені статистичні методи. Вони зосереджуються на практичних даних і дають змогу оцінити ступінь випадковості згенерованих чисел;

- *теоретичні*, які ґрунтуються на математичних моделях і використовуються для аналізу характеристик генератора з погляду статистичних властивостей. У цьому випадку оцінка здійснюється на основі теоретичних уявлень про ймовірні результати роботи алгоритму;

- *прикладні*, які, на відміну від попередніх, оцінюють ефективність генератора в реальних програмних застосунках. Вони проводяться на основі вже відомих аналітичних результатів, які можуть бути обчислені для конкретних сценаріїв.

Машинне навчання є однією з технологій, що розвиваються найбільш динамічно, і щодня з'являються нові інструменти та програми. Методи машинного навчання знаходять застосування у

різноманітних сферах та для розв'язання актуальних задач [17- 19]. Його популярність зростає завдяки здатності ефективно обробляти великі обсяги даних і робити точні прогнози на їх основі. Однією з головних причин цього є доступність величезних обсягів даних, оскільки розвиток інтернету та поширення мобільних технологій збільшили доступ до них. Алгоритми машинного навчання можуть аналізувати ці дані для виявлення закономірностей і створення прогнозів, які мають численні практичні застосування [20- 21]. Крім того, розвиток хмарних обчислень і використання графічних процесорів дають змогу значно підвищити швидкість і точність обробки даних. Завдяки цим технологіям машинне навчання стає потужним інструментом для вирішення складних завдань у різних галузях.

Розвиток машинного навчання зумовлений постійним удосконаленням алгоритмів і моделей, що забезпечує більш точні прогнози та складніші рішення. Важливим чинником є також попит на автоматизацію різних процесів у бізнесі, де машинне навчання допомагає оптимізувати завдання, як-от обробка даних чи обслуговування клієнтів. Крім того, доступність інструментів і бібліотек з відкритим кодом стимулює розвиток цієї технології, даючи змогу розробникам легко впроваджувати інноваційні рішення. Очікується, що в майбутньому машинне навчання стане ще більш значущим у різних сферах, зміцнюючи свою роль у технологічних і бізнес-інноваціях.

Постановка задачі

Хоча псевдовипадкові числа відіграють ключову роль у машинному навчанні, їхній вплив на результати алгоритмів досліджувався недостатньо. Отже, важливо глибше вивчити цей вплив та виявити можливі приховані ризики й обмеження, які вони можуть створювати.

Мета дослідження полягає у всебічному аналізі ролі генераторів псевдовипадкових чисел у сучасному машинному навчанні. Особливу увагу необхідно приділити впливу якості ГПВЧ на продуктивність та відтворюваність результатів навчання моделей, а також викликам, пов'язаним із передбачуваністю, енергоефективністю та масштабованістю генераторів у різних умовах. Також необхідно визначити основні проблеми та виклики застосування ГПВЧ у машинному навчанні та визначити рекомендації щодо оптимізації такого їхнього використання.

Роль генераторів псевдовипадкових чисел у машинному навчанні

Випадкові числа відіграють важливу роль у багатьох задачах машинного навчання (ML). Вони необхідні для реалізації стохастичних методів оптимізації, вибірки даних з розподілів генеративних моделей або для початкової ініціалізації параметрів нейронних мереж. Початкова ініціалізація параметрів мережі є критично важливою, оскільки визначає стартову точку для оптимізації, що може значною мірою вплинути на час, необхідний для навчання моделі, та її кінцеву ефективність. Правильно підібрані випадкові числа можуть допомогти уникнути проблем, як-от застрягання в локальних мінімумах, і забезпечити більш швидке та стабільне навчання. З розвитком алгоритмів машинного навчання роль псевдовипадкових чисел стає дедалі важливішою для досягнення високих результатів. Проте, попри зростаючу зацікавленість, важливими викликами залишаються проблеми відтворюваності результатів і високого енергоспоживання. Відтворюваність є ключовим елементом для забезпечення достовірності наукових досліджень і можливості перевірки їх результатів, тоді як енергоефективність акцентує увагу на раціональному використанні глобальних ресурсів. Особливо це важливо у зв'язку з дедалі більшою складністю моделей, які потребують значних обчислювальних потужностей. Удосконалення алгоритмів і підходів, спрямованих на підвищення ефективності, є критично важливим для сталого розвитку цієї галузі. Тому надійне джерело високоякісних випадкових чисел є важливим інструментом не лише в галузі штучного інтелекту, а й у багатьох інших сферах природничих наук, таких як методи шифрування та моделювання [22].

Штучні нейронні мережі навчаються за допомогою стохастичних методів оптимізації, зокрема стохастичного градієнтного спуску (SGD). Цей алгоритм використовує елементи випадко-

вості для ефективного знаходження найбільш вдалого набору вагових коефіцієнтів, які забезпечують оптимальне відображення між вхідними та вихідними даними. Завдяки стохастичному характеру алгоритму кожен запуск процесу навчання на одному і тому самому наборі даних може призводити до формування моделі з дещо різними характеристиками, що залежить від початкової ініціалізації ваг та порядку обробки мінібатчів.

Ця випадковість, попри її вплив на результати, є важливим фактором, який дозволяє уникати локальних мінімумів та покращує здатність моделі до генералізації. Використання відтворюваних початкових умов (seed) дає змогу зменшити варіативність між різними запусками та забезпечити контрольованість навчального процесу.

Ефективність процесу навчання нейронних мереж значною мірою визначає витрати обчислювальних ресурсів, їх розподіл та час, необхідний для створення моделі в межах машинного навчання. Одним із ключових факторів, що впливають на результативність навчання, є оптимізатор, чия здатність належно навчити модель і забезпечити її високу продуктивність безпосередньо залежить від початкових значень ваг мережі. Правильно вибрані початкові ваги можуть значно пришвидшити процес навчання, зменшити ймовірність застрягання в локальних мінімумах і підвищити якість кінцевих результатів. Крім того, оптимізація параметрів навчання, таких як швидкість навчання, також може істотно покращити ефективність процесу. Схеми ініціалізації ваги моделі використовують генератори псевдовипадкових чисел як джерело випадковості [23].

Сучасні дослідники машинного навчання (ML) переважно використовують високорівневі мови програмування та фреймворки для проведення своїх досліджень [24]. Python залишається основною мовою для розробки у сфері машинного навчання, що сприяло популярності фреймворків, таких як PyTorch і TensorFlow, часто у поєднанні з бібліотекою NumPy. Попри це, існує обмежена кількість досліджень, присвячених якості генераторів псевдовипадкових чисел у контексті машинного навчання. У Python стандартним ГПВЧ є Mersenne Twister (MT), тоді як у TensorFlow використовується Philox, із додатковою підтримкою Threefry з того самого класу криптографічно захищених генераторів. PyTorch пропонує схожі можливості, тоді як NumPy забезпечує більшу гнучкість завдяки доступу до кількох PRNG, серед яких за замовчуванням використовується PCG.

Зростання глибокого навчання та впровадження складних моделей у машинному навчанні потребує потужних обчислювальних ресурсів для обробки великих обсягів даних. Виробники апаратних прискорювачів активно працюють над покращенням продуктивності, пропонуючи рішення з високою швидкістю обчислень. Продуктивність, вимірювана часом обробки або швидкістю виконання операцій, визначає можливість тренувати масштабні моделі та повторювати експерименти на етапах досліджень. Оптимізація алгоритмів і використання сучасного апаратного забезпечення дає змогу скоротити час навчання, проте проблема енергоспоживання стає дедалі актуальнішою, особливо з урахуванням екологічної стійкості.

Високе енергоспоживання підвищує операційні витрати та негативно впливає на довкілля, сприяючи збільшенню вуглецевого сліду в центрах обробки даних. Тому критично важливо оптимізувати як продуктивність, так і енергоефективність обчислень, включно з генераторами псевдовипадкових чисел. Ефективні ГПВЧ здатні значно пришвидшити процеси ініціалізації, перетасовки та інші стохастичні операції, що використовуються у машинному навчанні, одночасно знижуючи як час виконання, так і споживання енергії. Це робить їх важливим елементом для підвищення ефективності ML-процесів у контексті сталого розвитку.

Відтворюваність у машинному навчанні відображає здатність точно відтворити модель, забезпечуючи узгодженість точності та прозорість її роботи. Однак через використання псевдовипадкових чисел під час створення та навчання моделі досягнення повної відтворюваності часто стає складним завданням. Цей аспект є критичним, оскільки навіть незначні варіації у випадкових компонентах можуть вплинути на кінцеву структуру та продуктивність моделі. Це важливий чинник для точного налаштування моделей, оцінки їхньої ефективності, перевірки достовірності результатів і забезпечення надійності в реальних сценаріях. Значну роль у досягненні відтворю-

ваності відіграють генератори псевдовипадкових чисел, оскільки процеси, такі як поділ даних та ініціалізація параметрів, часто залежать від псевдовипадкових послідовностей. У разі недосконалої роботи ГПВЧ навіть незначні відхилення можуть накопичуватися у процесі навчання, спричиняючи значні розбіжності в результатах.

Відтворюваність важлива не лише для окремих експериментів, а й для наукової спільноти загалом, адже вона забезпечує основу для подальших досліджень. Достовірні результати підвищують довіру до висновків, сприяють розробці стабільних технологій і пришвидшують реалізацію нових ідей у практичних застосуваннях. Крім того, вона знижує ризик дублювання роботи та сприяє ефективнішому використанню ресурсів, що особливо важливо в умовах сучасних швидких темпів розвитку науки й технологій.

Для підвищення довіри до моделей і забезпечення їхньої відтворюваності важливо контролювати генерацію випадкових чисел, зокрема шляхом фіксації початкових умов (seed) для генераторів псевдовипадкових чисел. Це дає змогу мінімізувати варіативність між навчальними запусками та полегшує точне відтворення результатів у контексті як академічних досліджень, так і практичного застосування.

Результати дослідження

Машинне навчання значною мірою базується на статистичних методах, і використання випадкових чисел є ключовим для реалізації багатьох етапів обробки даних та процесів навчання моделей. Від початкової ініціалізації параметрів до поділу даних на тренувальні та тестові набори – випадковість сприяє підвищенню точності та здатності алгоритмів до узагальнення. Крім того, генерація випадкових чисел забезпечує гнучкість і варіативність, необхідні для налаштування та оптимізації сучасних моделей. У прикладному машинному навчанні існує безліч джерел, які спричиняють випадковість. Її застосовують як засіб для підвищення стійкості алгоритмів, що в результаті сприяє створенню більш точних моделей і покращенню прогнозів. Випадковість дає змогу уникати перенавчання, забезпечуючи різноманітність у вибірках даних і процесах оптимізації. Це допомагає моделям краще адаптуватися до складних завдань і працювати ефективніше в умовах невизначеності.

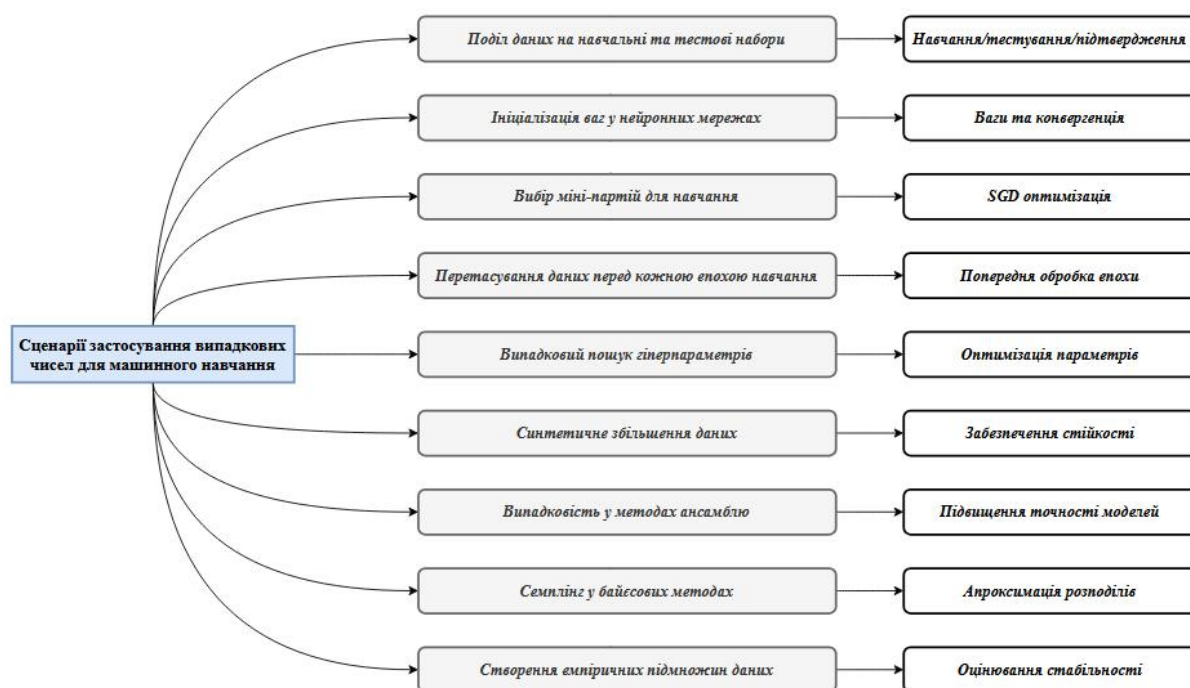


Рис. 1. Найпоширеніші сценарії застосування псевдовипадкових чисел для машинного навчання

Ми визначили **найпоширеніші сценарії** застосування псевдовипадкових чисел для машинного навчання (рис. 1):

1. Поділ даних на навчальні та тестові набори.

Одним із ключових кроків у процесі оцінки ефективності алгоритмів машинного навчання є поділ набору даних на навчальні та тестові підмножини. Основна мета цього етапу – створити модель, здатну узагальнювати знання на нові, невідомі дані. Для досягнення цього зразки даних розділяються на два або більше непересічних наборів: навчальний і тестовий.

Навчальний набір використовується для корекції параметрів моделі шляхом ітеративного налаштування, тоді як тестовий набір дає змогу об'єктивно оцінити продуктивність моделі. Використання випадковості під час цього розподілу зменшує ймовірність упередженості та гарантує репрезентативність вибірки, забезпечуючи точніше оцінювання здатності моделі до узагальнення.

2. Ініціалізація ваг у нейронних мережах.

Нейронні мережі потребують випадкової ініціалізації ваг, що є критичним для початкового етапу навчання. Загальною практикою є ініціалізація коефіцієнтів малими випадковими числами перед навчанням моделі, щоб зменшити можливість стагнації моделі [25]. Правильна ініціалізація запобігає проблемам, пов'язаним із симетрією вузлів, і сприяє швидшій збіжності оптимізаційного процесу. Якби всі вагові коефіцієнти були рівними, алгоритм зворотного поширення не мав би стимулів для різного оновлення кожного нейрона, оскільки відсутність варіативності у значеннях ваг призводила б до однакових змін для всіх елементів моделі. Це обмежує здатність моделі адаптуватися та ефективно вивчати різноманітні патерни в даних. Отже, на етапі ініціалізації моделі ваги часто вибираються випадковим чином для створення початкової асиметрії, що дозволяє алгоритму оптимізації розрізняти важливість різних нейронів під час процесу навчання. Такий підхід забезпечує кращу збіжність моделі та сприяє ефективному навчанню на складних даних.

Псевдовипадкові числа дають змогу розподілити ваги в межах діапазонів, які забезпечують належну стартову позицію для ефективного навчання. Невдалий вибір початкових ваг може значно уповільнити навчання або навіть зробити модель незбіжною.

3. Вибір мініпартій для навчання.

Застосування всього навчального набору даних для одноразового оновлення параметрів моделі є не лише обчислювально неефективним, але й з погляду практичності часто невиправданим. З цієї причини навчальний набір даних розділяється на мініпартії фіксованого розміру, що дає змогу здійснювати оновлення параметрів моделі поступово, знижуючи обсяг обчислень та прискорюючи процес навчання. Завантажувач даних відповідає за формування цих мініпакетів, при цьому часто включає випадкове перемішування даних для уникнення упередженості, яка може виникнути через порядок їх подачі (наприклад, у разі, коли дані зібрані у хронологічній послідовності). Стохастичні алгоритми оптимізації, такі як стохастичний градієнтний спуск (SGD) [26], базуються на використанні мініпартій даних для оновлення параметрів моделі. Випадковий вибір мініпартій не тільки підвищує ефективність навчання, а й зменшує ризик переобладнання, сприяє загальній узагальнювальній здатності моделі та покращує стабільність оптимізації. Крім того, це дозволяє алгоритму уникати локальних мінімумів і досягати кращих результатів.

4. Перетасування даних перед кожною епохою навчання.

Перетасування порядку даних у навчальному наборі перед початком кожної епохи є стандартною практикою для багатьох алгоритмів машинного навчання [27]. Це дає змогу уникнути потенційного впливу послідовності даних на результати навчання, що може стати джерелом систематичної похибки. Випадкове перетасування забезпечує більш рівномірний розподіл зразків, сприяючи стабільності та ефективності навчального процесу, особливо у випадках, коли дані містять певні впорядковані шаблони або групування.

5. *Випадковий пошук гіперпараметрів.*

У процесі налаштування моделей машинного навчання важливу роль відіграє вибір оптимальних гіперпараметрів. Одним із методів їхнього підбору є випадковий пошук (Random Search), під час якого значення гіперпараметрів вибираються випадковим чином із визначених діапазонів. Цей підхід дає змогу досліджувати ширший простір параметрів, часто виявляючи комбінації, які могли б залишитися поза увагою за використання систематичного підходу, як-от ґратчастий пошук (Grid Search) [28].

6. *Синтетичне збільшення даних.*

У задачах, де доступні набори даних, є обмеженими, синтетичне збільшення даних (Data Augmentation) використовується для створення нових зразків шляхом внесення випадкових змін до існуючих даних. Наприклад, у комп'ютерному зорі це може спричиняти випадкове обертання, масштабування, зсуви або зміни кольорової гами зображень. Такі операції збагачують навчальний набір і допомагають моделям стати більш стійкими до варіацій у реальних даних.

7. *Випадковість у методах ансамблю.*

Ансамблеві методи, такі як випадкові ліси (Random Forests) [29], активно використовують випадковість для підвищення точності моделей. Наприклад, під час побудови кожного дерева в лісі випадково вибирається підмножина характеристик або зразків даних. Це сприяє створенню більш різноманітних дерев, що знижує ризик переобладнання та підвищує загальну стабільність і продуктивність ансамблю.

Випадкові ліси є методом ансамблевого навчання, який широко застосовується для задач класифікації та регресії. Цей метод базується на сукупності дерев рішень, які незалежно здійснюють передбачення класу або значення для вхідних даних. Остаточний прогноз формується на основі агрегованих результатів дерев, зазвичай шляхом голосування більшістю у випадку класифікації або усереднення для задач регресії.

Випадковість відіграє ключову роль у структурі та навчанні випадкових лісів. По-перше, випадкові числа використовуються для вибору підмножини характеристик під час розгляду можливих поділів даних у кожному вузлі дерева. По-друге, випадковість застосовується для створення підмножин навчальних даних, що забезпечує різноманітність дерев у лісі та сприяє зменшенню варіативності та переобладнання моделі. Такий підхід робить випадкові ліси надійним і універсальним інструментом, здатним ефективно працювати з високовимірними даними та уникати переобладнання навіть за значної кількості дерев.

8. *Семплінг у байєсових методах.*

У байєсовому машинному навчанні випадкові числа відіграють ключову роль у методах семплінгу, таких як ланцюги Маркова Монте-Карло (MCMC). Ці методи використовують випадковість для ітеративного дослідження розподілів і обчислення ймовірностей, що дає змогу ефективно працювати з високовимірними просторами параметрів.

9. *Створення емпіричних підмножин даних.*

Такі методи, як бутстрепінг (Bootstrap), використовують випадкові числа для створення підмножин даних із повтореннями. Це дає змогу оцінювати стабільність моделей, зменшувати похибки оцінок і покращувати якість передбачень шляхом комбінування результатів із декількох підвбірок.

Отже, випадкові числа забезпечують основу для багатьох методів і процесів у машинному навчанні, даючи змогу створювати моделі, які є стійкими, точними та адаптивними до різноманітних задач і умов.

Варто зазначити, що для задач машинного навчання рівень випадковості, що вимагається, значно нижчий, порівняно з тим, який необхідний для криптографічних алгоритмів. У машинному навчанні випадковість використовується для забезпечення ефективної ініціалізації параметрів моделей та оптимізації, однак не потребує такої високої непередбачуваності, як у випадку з

криптографічними системами. Ключовою метою є здатність моделі навчатися та адаптуватися до даних, зберігаючи стабільність і забезпечуючи узагальнення на нові зразки. Саме тому не обов'язково використовувати криптографічно стійкі генератори псевдовипадкових чисел.

Однак у машинному навчанні важливо використовувати достатню кількість випадкових елементів для уникнення переобладнання і покращення здатності моделі до генералізації, особливо під час роботи з великими та складними даними. Для цього необхідно, щоб генератори псевдовипадкових послідовностей, які використовуються в машинному навчанні, забезпечували якомога більший період повторення.

Також не висуваються особливі вимоги щодо закону розподілу, за яким мають бути згенеровані псевдовипадкові числа. Зазвичай з рівномірно розподіленої послідовності можна отримати послідовності з іншими законами розподілу. У контексті машинного навчання це дає змогу створювати різноманітні варіації даних для навчання моделей, що сприяє підвищенню їхньої узагальнювальної здатності. Використання таких трансформацій допомагає забезпечити більшу різноманітність у навчальних даних, що може покращити стабільність та точність моделей, особливо у задачах, де присутня велика варіативність або шум.

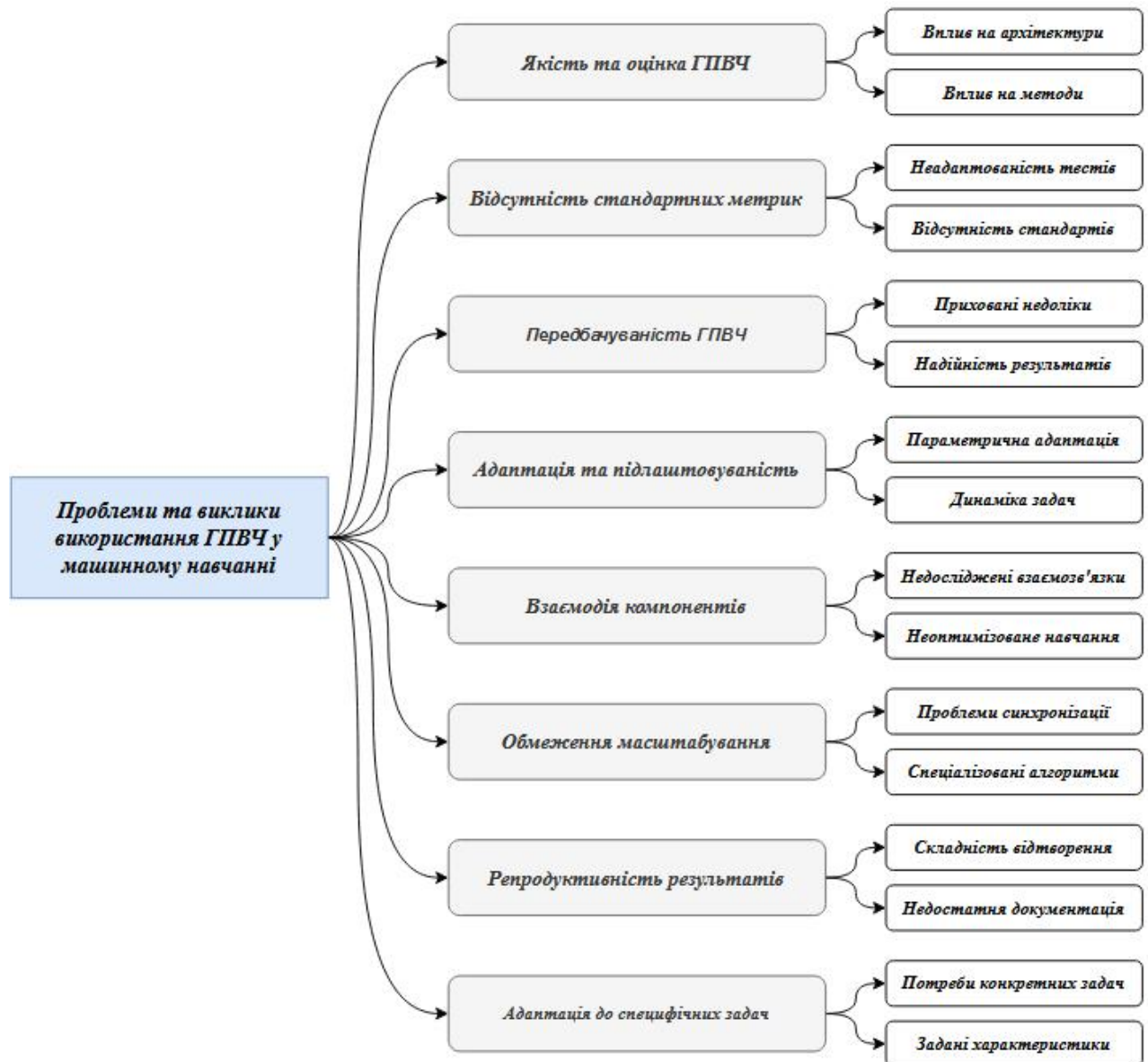


Рис. 2. Основні проблеми та виклики застосування ГПВЧ для машинного навчання

У сфері використання генераторів псевдовипадкових чисел (ГПВЧ) для машинного навчання також виділили низку важливих прогалин, які потребують подальших досліджень (рис. 2). Нижче наведено **основні аспекти**, які варто розглянути:

1. Вплив якості ГПВЧ на моделі машинного навчання.

Існує обмежена кількість досліджень, що вивчають вплив різних типів генераторів і їх якості на продуктивність різних моделей. Хоча якість ГПВЧ відома як критичний фактор для ефективного навчання, її вплив на різноманітні архітектури, такі як нейронні мережі, дерева рішень або методи кластеризації та регресії, залишається недостатньо дослідженим. Глибше розуміння цих взаємозв'язків може допомогти розробити рекомендації щодо вибору генераторів для конкретних застосувань.

2. Відсутність стандартних метрик оцінки якості ГПВЧ у машинному навчанні.

Наявні статистичні тести для оцінки якості ГПВЧ зазвичай не адаптовані для специфічних потреб машинного навчання. Наприклад, немає єдиного стандарту для аналізу довгострокових ефектів низької якості ГПВЧ на стабільність результатів навчання. Це особливо важливо для задач із високою чутливістю до ініціалізації, таких як нейронні мережі. Запровадження нових метрик та методів тестування може сприяти більш точній оцінці впливу ГПВЧ на результати.

3. Проблема періодичності та передбачуваності ГПВЧ.

Періодичність і передбачуваність роботи деяких ГПВЧ залишаються важливими проблемами. Ці властивості можуть створювати приховані недоліки, які проявляються лише за специфічних умов. Розробка методів для оцінки та виявлення таких обмежень допоможе підвищити надійність результатів навчання.

4. Незначна увага до адаптивних ГПВЧ.

Більшість досліджень зосереджені на статичних генераторах, тоді як адаптивні ГПВЧ, які можуть змінювати свої параметри під час навчання, здатні забезпечити вищу якість результатів. Такі генератори можуть враховувати специфіку задачі, підлаштовуючись під її динаміку, що робить їх перспективним напрямом для досліджень.

5. Взаємодія між ГПВЧ та іншими компонентами моделей.

Залишається недостатньо вивченою взаємодія між ГПВЧ і такими параметрами моделей, як швидкість навчання, розмір мініпартій або особливості архітектури. Розуміння цих взаємозв'язків дозволить оптимізувати процес навчання та підвищити ефективність алгоритмів.

6. Обмеження масштабування для великих даних і розподілених систем.

Використання стандартних ГПВЧ у великих наборах даних та розподілених обчислювальних середовищах може призводити до синхронізаційних проблем. Розробка спеціалізованих алгоритмів для таких сценаріїв допоможе забезпечити точність та узгодженість результатів.

7. Репродуктивність результатів досліджень.

Однією з основних проблем є складність відтворення результатів експериментів через недостатню інформацію про використані ГПВЧ, їх параметри та експериментальні умови. Запровадження стандартів документування та звітності сприятиме підвищенню прозорості та наукової цінності досліджень.

8. Адаптація ГПВЧ до специфічних задач.

Більшість ГПВЧ орієнтовані на створення випадкових чисел із рівномірним розподілом, що не завжди відповідає потребам конкретних задач, таких як байесовий аналіз або ансамблеві методи. Розробка адаптованих генераторів із іншими заданими характеристиками розподілу є перспективним напрямом для подальшого вдосконалення методів машинного навчання.

Поглиблення досліджень за цими напрямками сприятиме створенню більш надійних алгоритмів і покращенню якості рішень у галузі машинного навчання.

Також варто згадати й інший бік цієї взаємодії, коли нейронні мережі використовуються для виявлення послідовностей, згенерованих генераторами псевдовипадкових чисел, а також для атак на ці генератори. Такий підхід дозволяє нейронним мережам знаходити закономірності у здавалося

б випадкових числових послідовностях і використовувати їх для зламування генераторів. Звичайний генератор псевдовипадкових чисел вразливий до атак машинного навчання, оскільки для генерації випадкових чисел використовуються алгоритми. Останнім часом глибоке навчання (DL) продемонструвало значний прогрес у вирішенні складних завдань, що призвело до припущень серед дослідників про його потенціал у виявленні закономірностей у здавалося б випадкових об'єктах, які раніше були недоступні для традиційних підходів. Ця здатність аналізувати складні дані й знаходити приховані взаємозв'язки відкриває нові можливості для досліджень у багатьох сферах. З розвитком DL перспективи розширюються, і це може змінити підхід до задач, які раніше вважалися нерозв'язними через їхню складність.

Виявлення слабких місць у генераторах псевдовипадкових чисел може мати важливі наслідки для безпеки, оскільки ці генератори часто використовуються в криптографії та інших чутливих системах. Застосування нейронних мереж в таких задачах відкриває нові можливості для розробки більш безпечних алгоритмів, але й створює нові виклики у сфері захисту від подібних атак [30–33].

Висновки

Результати дослідження підтвердили критичну роль ГПВЧ у забезпеченні ефективності та стабільності моделей машинного навчання. Використання якісних генераторів сприяє підвищенню точності алгоритмів та їх здатності до узагальнення, зокрема в задачах ініціалізації нейронних мереж та стохастичної оптимізації. Проте виявлено низку проблем, зокрема низьку адаптивність існуючих генераторів до специфічних задач, обмеження періодичності та передбачуваності, а також виклики, пов'язані з енергоефективністю. Важливим напрямом є вдосконалення ГПВЧ шляхом їх адаптації до задач із високою чутливістю до випадкових компонентів. Перспективним є також використання глибокого навчання для аналізу та виявлення слабких місць у генераторах. Запропоновані рекомендації сприятимуть створенню більш надійних і продуктивних ML-систем, що мають велике значення для наукових досліджень та практичних застосувань.

Список літератури

1. Gnatyuk S., Burmak Y., Berdibayev R., Aleksander M., Ospanova D. Метод побудови генераторів псевдовипадкових послідовностей для криптографічних застосувань у 5G мережах. Електронне фахове наукове видання "Кібербезпека: освіта, наука, техніка". Червень 2021. Вип. 4, Вип. 12. С. 151- 62. DOI: 10.28925/2663-4023.2021.12.151162.
2. Гарасимчук О. І., Максимович В. М. Генератори псевдовипадкових чисел, їх застосування, класифікація, основні методи побудови і оцінка якості. Захист інформації. 2002. 5(3 (16)). С. 29- 36.
3. Хомік М. А., Гарасимчук О. І. Застосування генераторів псевдовипадкових чисел та послідовностей в кібербезпеці, методи їх побудови та оцінки якості. Захист інформації. Липень-вересень 2023. Т. 25. № 3. DOI: 10.18372/2410-7840.25.17940.
4. Поперешняк С. В. Застосування генератора псевдовипадкових чисел для підвищення ефективності технології smart dust в управлінні розумним будинком. Телекомунікаційні та інформаційні технології. 2022. № 4 (77). DOI: 10.31673/2412-4338.2022.045362.
5. Orúe A. B., Hernández Encinas L., Fernández V., Montoya F. A Review of Cryptographically Secure PRNGs in Constrained Devices for the IoT. In Proceedings of the SOCO 2017, ICEUTE 2017, CISIS 2017: International Joint Conference SOCO'17-CISIS'17-ICEUTE'17 León, Spain, 6–8 September 2017. DOI: 10.1007/978-3-319-67180-2_65.
6. Maksymovych V., Shabatura M., Harasymchuk O., Shevchuk R., Sawicki P., Zajac T. Combined Pseudo-Random Sequence Generator for Cybersecurity. Sensors (Basel). 2022. 22. 9700. DOI: 10.3390/s22249700.
7. Maksymovych V., Shabatura M., Harasymchuk O., Karpinski M., Jancarczyk D., Sawicki P. Development of Additive Fibonacci Generators with Improved Characteristics for Cybersecurity Needs. Appl. Sci. (Basel). 2022. 12. 1519. DOI: 10.3390/app12031519.
8. Almaraz Luengo E. A brief and understandable guide to pseudo-random number generators and specific models for security. Statistic Surveys. 2022. 16. 137- 181. DOI: 10.1214/22-SS136.

9. Горбенко С. І., Шапочка Н. В., Грінченко Т. О., Нейванов А. В., Мордвінов Р. І. Методи та засоби генерування псевдовипадкових послідовностей. 2011.
10. Горбенко І. Д., Шапочка Н. В., Козулін О. О. Обґрунтування вимог до генераторів випадкових бітів згідно ISO/IEC 18031. *Радіоелектронні і комп'ютерні системи*. 2009. 6. С. 94- 97.
11. Knuth Donald E. *The Art of Computer Programming*. 3rd ed., Addison Wesley, 1997.
12. A Statistical Test Suite for Random and Pseudorandom Number Generators for Cryptographic Applications. Andrew Rukhin, Juan Soto, James Nechvatal, Miles Smid, Elaine Barker, Stefan Leigh, Mark Levenson, Mark Vangel, David Banks, Alan Heckert, James Dray, San Vo. NIST Special Publication 800-22, Revision 1a, April, 2010.
13. Соколовська Г. В. Статистичний аналіз генераторів псевдовипадкової послідовності у програмних середовищах Matlab та Mathcad. Моделювання та інформаційні технології. 2013. Вип. 66. С. 26–30.
14. Alani M. M. Testing randomness in ciphertext of block-ciphers using DieHard tests. *International Journal of Computer Science and Network Security (IJCSNS)*. 2010. 10(4). 53–57.
15. NIST SP 800-22 Version 1a. A Statistical Test Suite for Random and Pseudorandom Number Generators for Cryptographic Applications; NIST: Gaithersburg, MD, USA, 2010, p. 131. URL: <https://nvlpubs.nist.gov/nistpubs/Legacy/SP/nistspecialpublication800-22r1a.pdf> (accessed on 20 April 2023).
16. Min Lequan et al. Analysis of FIPS 140-2 Test and Chaos-Based Pseudorandom Number Generator. 2013.
17. Susan Athey. *The Impact of Machine Learning on Economics*. In *The economics of artificial intelligence*. University of Chicago Press, 2019. 507–552.
18. Партика А., Михайлова О., Шпак С. Виявлення, аналіз та захист конфіденційних даних за допомогою технології машинного навчання сервісу AMAZON MACIE. Електронне фахове наукове видання “Кібербезпека: освіта, наука, техніка”. 2024. 3(23). 132–144. DOI: 10.28925/2663-4023.2024.23.132144.
19. Kovalchuk O., Karpinski M., Banakh S., Kasianchuk M., Shevchuk R., Zagorodna N. Prediction Machine Learning Models on Propensity Convicts to Criminal Recidivism. *Information*. 2023. 14. 161. DOI: 10.3390/info14030161.
20. Партика А., Гарасимчук О., Нємкова О., Совин Я., Дудикевич В. Розроблення методу дослідження кіберзлочинів за типом вірусів-вимагачів з використанням моделей штучного інтелекту в системі менеджменту інформаційної безпеки критичної інфраструктури. *Social Development and Security*. 2024. 14(2). 52- 63. DOI: 10.33445/sds.2024.14.2.6.
21. Deineka O., Harasymchuk O., Partyka A., Kozachok V. Information classification framework according to SOC 2 Type II. *Cybersecurity Providing in Information and Telecommunication Systems II*. 2024. 3826. Pp. 182- 189. ISSN 1613-0073.
22. Hanin Boris, Rolnick David. How to Start Training: The Effect of Initialization and Architecture. 2018. DOI: 10.48550/arXiv.1803.01719.
23. Miranskyy A., Sorrenti A., Thakar V. On Using Quasirandom Sequences in Machine Learning for Model Weight Initialization, 2024. DOI: 10.48550/arXiv.2408.02654.
24. Antunes B., Hill D. R. C. Random Numbers for Machine Learning: A Comparative Study of Reproducibility and Energy Consumption. *Journal of Data Science and Intelligent Systems*. 2024. DOI: 10.47852/ bonview JDSIS42024012.
25. Jason Brownlee. *How to Choose an Activation Function for Deep Learning*, 2021.
26. Catapang J. K. Optimizing Speed and Accuracy Trade-off in Machine Learning Models via Stochastic Gradient Descent Approximation. 2022 9th International Conference on Soft Computing & Machine Intelligence (ISCM), Toronto, ON, Canada, 2022. Pp. 124- 128. DOI: 10.1109/ISCM56532.2022.10068476.
27. Wan K., Tuninetti D., Ji M., Piantanida P. Fundamental Limits of Distributed Data Shuffling. 2018 56th Annual Allerton Conference on Communication, Control, and Computing (Allerton), Monticello, IL, USA, 2018. Pp. 662- 669. DOI: 10.1109/ALLERTON.2018.8635882.
28. Chang C. Y., Lee C. H., Chiang K. N. Using Grid Search Methods and Parallel Computing to Reduce AI Training Time for Reliability Lifetime Prediction of Wafer-Level Packaging. 2023 24th International Conference on Thermal, Mechanical and Multi-Physics Simulation and Experiments in Microelectronics and Microsystems (EuroSimE), Graz, Austria, 2023. Pp. 1- 5. DOI: 10.1109/EuroSimE56861.2023.10100751.

29. Shingade S. D., Mudhalwadkar R. P., Masal K. M. Random Forest Machine Learning Classifier for Seed Recommendation. 2022 International Conference on Edge Computing and Applications (ICECAA), Tamilnadu, India, 2022. Pp. 1385- 1390. DOI: 10.1109/ICECAA55415.2022.9936120.
30. Proskurin D. et al. Predicting pseudo-random number generator output with sequential analysis. CSDP-2024: Cyber Security and Data Protection, June 30, 2024, Lviv, Ukraine. Pp. 42- 57.
31. Proskurin D. et al. Hybrid RNN-CNN-based model for PRNG identification, CQPC-2024: Classic, Quantum, and Post-Quantum Cryptography, August 6, 2024, Kyiv, Ukraine. Pp. 47–53.
32. Amigo G., Dong L., Marks Ii R. J. Forecasting Pseudo Random Numbers Using Deep Learning. 2021 15th International Conference on Signal Processing and Communication Systems (ICSPCS), Sydney, Australia, 2021. Pp. 1- 7. DOI: 10.1109/ICSPCS53099.2021.9660301.
33. Лесик В. О., Дорошенко А. Ю. Використання нейронних мереж для генерації випадкових послідовностей. Проблеми програмування. 2024. № 2- 3. С. 280–287. DOI: 10.15407/pp2024.02-03.280.

APPLICATION OF PSEUDORANDOM NUMBER GENERATORS IN MACHINE LEARNING

O. I. Harasymchuk¹, S. S. Voytusik², O. A. Tsebak²

Lviv Polytechnic National University,

¹Department of Information Protection,

²Department of Information Technology Security

© Harasymchuk O. I., Voytusik S. S., Tsebak O. A., 2025

The paper analyzes the role of pseudorandom number generators in ensuring the efficiency and stability of machine learning algorithms. The key areas and forms of such application are identified.

Particular attention is paid to the impact of the quality of PRNGs on the performance of models, in particular in the tasks of initializing neural networks, synthetic data augmentation, stochastic optimization, and reproducibility of results. The problems of periodicity, predictability, and low adaptability of generators to specific ML tasks are considered. Modern approaches to assessing the quality of PRNGs, including empirical, theoretical, and applied tests, are analyzed. The study also highlights the prospects of using deep learning to analyze PRNGs and identify patterns in pseudorandom sequences.

Recommendations for improving the efficiency of machine learning systems by optimizing the use of PRNGs are proposed. The results of the study contribute to a deeper understanding of the role of PRNGs in modern machine learning and identify new directions for improving model performance.

Keywords: pseudorandom number generators, machine learning, neural network initialization, stochastic optimization, reproducibility.