КІБЕРБЕЗПЕКА: освіта, наука, техніка
CYBERSECURITY: EDUCATION, SCIENCE, TECHNIQUE
№ 3 (31), 2025
ISSN 2663 - 4023
DOI 10.28925/2663-4023.2025.31.1087
УДК 004.56.53

Вербиненко Віталій Олегович
магістрант кафедри систем та технологій кібербезпеки
Державний університет інформаційно-комунікаційних технологій, Київ, Україна
ORCID: 0009-0004-2134-1987
vv.q@ukr.net

Зибін Сергій Вікторович
доктор технічних наук, професор,
професор кафедри систем та технологій кібербезпеки
Державний університет інформаційно-комунікаційних технологій, Київ, Україна
професор кафедри прикладних інформаційних систем
Київський національний університет імені Тараса Шевченка
ORCID: 0000-0002-2670-2823
zysv@ukr.net

МЕТОД ПІДВИЩЕННЯ ЕФЕКТИВНОСТІ ДЕТЕКТУВАННЯ ІНСАЙДЕРСЬКИХ ЗАГРОЗ ІЗ ЗАСТОСУВАННЯМ GAN-АУГМЕНТАЦІЇ

Анотація. У сучасних корпоративних інформаційних системах значну частку інцидентів інформаційної безпеки становлять інсайдерські загрози, що породжує нові вимоги до систем моніторингу та аналізу подій безпеки. На відміну від зовнішніх атак, інсайдерська активність маскується під звичайну роботу легітимних користувачів, а тому важко описується за допомогою класичних сигнатурних або периметрових механізмів захисту. Додатковою складністю є крайній дисбаланс класів у журналах подій. Кількість записів про типову щоденну активність у тисячі разів перевищує число зафіксованих інцидентів, що призводить до деградації якості роботи стандартних алгоритмів машинного навчання. У статті розроблений підхід до підвищення ефективності детектування інсайдерських загроз шляхом аугментації даних із використанням генеративно-змагальних мереж, зокрема архітектури Conditional Tabular GAN (CTGAN). Запропоновано процес підготовки поведінкових логів, який передбачає агрегацію багатоканальних подій до рівня "користувач–день", побудову вектору динамічних поведінкових ознак та статичного контексту, логарифмічну нормалізацію ознак із "важкими хвостами" та масштабування до діапазону [–1; 1], що забезпечує стабільне навчання генеративної моделі. CTGAN налаштовується на моделювання умовного розподілу табличних даних міноритарного класу (інсайдерських атак) з урахуванням контексту ролі користувача та підрозділу. Для кожної неперервної ознаки застосовується спеціалізована нормалізація, яка дозволяє коректно відтворювати мультимодальні розподіли, а для дискретних змінних – техніка Gumbel-Softmax, що робить можливим навчання за методом зворотного поширення похибки. Запропонований метод є перспективним для інтеграції до систем класу SIEM/UEBA та подальшого поєднання з методами пояснюваного штучного інтелекту.

Ключові слова: інсайдерські загрози; генеративно-змагальні мережі; аугментація даних; виявлення аномалій; інформаційна безпека.

ВСТУП

Традиційні підходи до забезпечення кібербезпеки, що ґрунтуються на захисті периметра та сигнатурному аналізі, виявляються недостатніми в умовах сучасних інформаційних систем, де значну частку інцидентів становлять інсайдерські загрози – дії співробітників, які мають легітимний доступ до інформаційних ресурсів, але

зловживають ним [1]. На відміну від класичних зовнішніх атак, інсайдерська активність маскується під нормальну роботу користувача, тому важко піддається формалізації у вигляді жорсткого набору сигнатур чи статичних правил. Інсайдер, маючи актуальні облікові дані та права доступу, може фактично обходити традиційні периметрові засоби захисту, маскуючи зловмисну активність під рутинну роботу. Прихований і динамічний характер поведінки внутрішніх зловмисників робить їх виявлення надзвичайно складним. Згідно з дослідженнями, значна частка кіберінцидентів припадає саме на інсайдерів [2].

Основним джерелом даних для виявлення інсайдерської активності в сучасних організаціях є журнали (логи) дій користувачів, які генеруються різними інформаційними системами. До таких логів належать, зокрема, записи із сервісів віддаленого доступу, систем контролю доступу до файлових ресурсів, веб-серверів, засобів автентифікації, корпоративних поштових та офісних платформ. Різні компоненти інфраструктури формують набори подій, які у сукупності утворюють детальний цифровий слід діяльності кожного співробітника. Аналіз поведінки користувачів на основі таких аудиторських даних (хост-логи, мережні логи, контекстна інформація про облікові записи тощо) становить основу більшості підходів до виявлення інсайдерів. Водночас робота з цими даними пов’язана з низкою істотних труднощів, подолання яких потребує використання спеціалізованих методів попередньої обробки, моделювання та аналізу [3].

Без належної трансформації та агрегування журналів до форми, яка, з одного боку, зберігає істотну поведінкову інформацію, а з іншого — зменшує надлишкову варіативність атрибутів, побудова ефективних методів виявлення аномалій або шкідливої активності практично неможлива. На практиці кількість потенційно корисних ознак, що можуть бути виділено з детальних логів, зростає майже експоненційно зі збагаченням даних. Наприклад, в одному з досліджень з використанням CERT-датасету для виявлення інсайдерів було автоматично сформовано близько 70 тисяч поведінкових ознак на одного користувача з журналів активності, після чого довелося застосувати метод головних компонент для зниження розмірності простору ознак. Цей приклад демонструє масштаб проблеми: надто велика кількість характеристик не лише ускладнює побудову моделей, а й загрожує надмірним перенавчанням.

Критичною проблемою у цій сфері є дисбаланс класів: кількість записів про типову щоденну активність у журналах подій у тисячі разів перевищує число зафіксованих інцидентів. У таких умовах класичні алгоритми машинного навчання схильні оптимізуватися під домінуючий "нормальний" клас, ігноруючи рідкісні аномалії. Це призводить до високих показників загальної точності при одночасно незадовільних значеннях повноти для цільового класу інсайдерів.

Перспективним шляхом подолання зазначених обмежень є застосування генеративного штучного інтелекту [4], зокрема генеративно-змагальних мереж (GAN), для синтезу додаткових прикладів аномальної активності. Такі моделі, навчаючись на реальних прикладах атак, здатні породжувати нові сценарії поведінки, що зберігають логічну структуру та статистичні властивості вихідних даних. Це дозволяє збалансувати тренувальні вибірки без штучного спотворення простору ознак і підвищити здатність детекторів до виявлення інсайдерських загроз.

Аналіз останніх досліджень і публікацій. Сучасні роботи розглядають GAN і їхні модифікації як потужний механізм моделювання складних розподілів даних, зокрема табличних і мережевих, характерних для задач виявлення інсайдерських загроз та навіть систем виявлення вторгнень.

Особливо перспективним може бути застосування GAN у ситуаціях, коли бракує даних з мітками або спостерігається значний дисбаланс класів. Виявлення інсайдерських загроз належить саме до таких задач, оскільки шкідливі дії зловмисників-інсайдерів трапляються дуже рідко порівняно з нормальними діями працівників, і відповідні вибірки є вкрай незбалансованими [5]. Генеративні моделі можуть допомогти розв'язати цю проблему двома шляхами: шляхом аугментації даних, тобто генерування додаткових штучних прикладів рідкісної шкідливої поведінки для поповнення навчальних наборів, або шляхом безпосереднього виявлення аномалій, коли сама GAN-модель слугує детектором нетипових дій [6].

У першому підході GAN використовується як інструмент синтезу даних. Наприклад, у [7] показано, що застосування WCGAN-GP для генерування додаткових сценаріїв атак дає змогу ефективно розширити датасет інсайдерських інцидентів і зменшити дисбаланс класів.

Інше сучасне дослідження [8] запропонувало GAN-архітектуру під назвою SPCAGAN, що поєднує генеративну модель з методом головних компонент. Це дає змогу навчатись на гетерогенних даних про дії користувачів.

Також GAN можна безпосередньо залучати до пошуку аномалій. У такому випадку генеративна мережа навчається лише на нормальних даних, після чого аномальні дії виявляються як такі, що погано відтворюються генератором або ж викликають велику невпевненість дискримінатора [9].

Існуючі підходи до виявлення інсайдерських загроз можна умовно поділити на правила-орієнтовані та даних-орієнтовані. У першому випадку використовуються експертно сформульовані політики доступу й евристичні правила, які описують явно заборонені дії користувачів. Такий підхід є зрозумілим і добре контрольованим, але погано масштабується та не здатний врахувати всі можливі варіації інсайдерської поведінки [3].

Даних-орієнтовані методи спираються на аналіз великих обсягів поведінкових логів за допомогою методів машинного навчання та інтелектуального аналізу даних [10, 11]. Вони включають побудову профілів нормальної активності користувачів, детектування відхилень, моделювання послідовностей подій тощо. Значна частина робіт фокусується на способах агрегування різнорідних логів, виборі інформативних поведінкових ознак, а також розробленні метрик ризику користувача.

Окремим напрямом є дослідження методів балансування класів для задач кібербезпеки. Найбільш поширені підходи ґрунтуються на скороченні вибірки мажоритарного класу або доповненню міноритарного класу. Серед останніх відомою є техніка SMOTE (Synthetic Minority Over-sampling Technique), яка виконує лінійну інтерполяцію між найближчими сусідами міноритарного класу, створюючи додаткові синтетичні точки. Попри популярність, SMOTE має суттєві обмеження при роботі з гетерогенними табличними даними, характерними для поведінкових логів.

З розвитком генеративного моделювання на основі GAN з'явилися роботи, присвячені аугментації табличних даних, у тому числі для задач виявлення аномалій. Архітектура CTGAN була запропонована як спеціалізоване рішення для змішаних (неперервних і категоріальних) ознак із мультимодальними розподілами, що робить її привабливою для застосування в контексті інсайдерських загроз. Втім, питання практичної ефективності GAN-аугментації у порівнянні з класичними методами балансування для реалістичних датасетів на кшталт CERT залишається актуальним і потребує поглиблених експериментальних досліджень.

Мета статті. Метою роботи є розроблення та експериментальна перевірка методу підвищення ефективності детектування інсайдерських загроз у корпоративних мережах шляхом аугментації даних за допомогою генеративно-змагальних мереж на табличних поведінкових логах. Для досягнення цієї мети необхідно:

– сформувати єдиний семантичний простір для різнорідних журналів користувацької активності;

– здійснити розробку схеми попередньої обробки й нормалізації даних, придатну для навчання GAN;

– застосувати й виконати налаштування архітектури CTGAN для генерації валідних зразків інсайдерської активності;

– виконати порівняння ефективності запропонованого підходу з базовим навчанням на незбалансованих даних та балансуванням методом SMOTE.

ТЕОРЕТИЧНІ ОСНОВИ ДОСЛІДЖЕННЯ

Інсайдерська загроза зазвичай визначається як зловмисна дія з боку поточного або колишнього співробітника, який має авторизований доступ до систем та зловживає ним. Складність виявлення інсайдера зумовлена тим, що його шкідливі дії майже не відрізняються від повсякденних легітимних операцій. Інсайдер, маючи актуальні облікові дані, може фактично обходити традиційні периметрові засоби захисту.

Основним джерелом даних для виявлення інсайдерської активності є журнали дій користувачів. У типових корпоративних журналах подій частка зловмисних інцидентів мізерна (часто менше ніж 1%), що є крайнім випадком задачі з незбалансованими даними. Стандартні підходи балансування, такі як SMOTE, базуються на лінійній інтерполяції між наявними прикладами міноритарного класу [12].

$$x_{new} = x_i + \lambda\left(x_{nn} - x_i\right), \lambda \sim U\left(0,1\right).$$

Застосування SMOTE до логів поведінкової активності має декілька суттєвих обмежень:

– Порушення семантичної цілісності. Вектори ознак у логах є гетерогенними (містять і неперервні, і дискретні змінні). Лінійна інтерполяція між вектором IT-адміністратора та вектором бухгалтера може створити варіант гібридного користувача, що не відповідає жодній реальній ролі.

– Ігнорування мультимодальності. Розподіли поведінкових метрик часто мають кілька локальних мод (наприклад, піки активності вранці та ввечері). SMOTE може генерувати зразки в областях низької ймовірності між модами, утворюючи шум.

– Втрата структурних залежностей. Логи подій мають внутрішню логіку й часову послідовність, яку прості геометричні методи аугментації даних часто порушують.

На відміну від цього, використання генеративно-змагальних мереж дозволяє навчитися повному розподілу даних і створювати синтетичні приклади, які зберігають логічну структуру та статистичні взаємозв’язки між ознаками.

Для подолання зазначених обмежень авторами пропонується використовувати генеративний підхід на основі архітектури CTGAN, який спеціально розроблений для моделювання складних табличних розподілів зі змішаними типами даних [13]. Розподіл кожної числової колонки C_j розглядається як суміш нормальних компонент і моделюється за допомогою варіаційної моделі суміші нормальних розподілів Гауса (VGM):

$$P(c_j) = \sum_{k=1}^{K} w_k N(c_j; \mu_k, \sigma_k),$$

де K – кількість мод (компонент);

w_k, μ_k, σ_k – вага, середнє та дисперсія k-ї моди відповідно.

У процесі попередньої обробки кожне значення v трансформується у вектор, що складається з:

– одиничного вектора α, який вказує на приналежність значення до конкретної моди (компоненти GMM);

– скаляра β, що представляє нормалізоване значення в межах цієї моди.

Такий підхід дозволяє нейромережі ефективно вивчати не лише діапазон значень, а й складну форму розподілу (наприклад, наявність важких хвостів у розподілі файлових операцій).

Генератор побудовано як глибоку нейронну мережу із залишковими зв'язками. На вхід подається конкатенація шумового вектору $z \sim N(0, I)$ та умовного вектору c. Умовний вектор формується шляхом об'єднання кодованого одиничним вектором класу (наприклад, інсайдер/норма) та маски, що вказує на конкретну категорію в одній з дискретних колонок.

Методика підготовки даних та інженерія ознак

Ефективність GAN-аугментації критично залежить від якості вхідних даних. У межах дослідження використано датасет CERT Insider Threat Dataset r4.2 [14], який містить багатоканальні журнали подій (Logon, Device, File, HTTP, Email). Методика передбачає інтеграцію цих гетерогенних потоків у єдиний семантичний простір.

Ключовим рішенням є вибір базової одиниці аналізу – агрегований інтервал типу "користувач–день". Такий підхід поєднує статистичну повноту опису з чутливістю до добових патернів. Для кожного об'єкта формується вектор динамічних поведінкових ознак $x \in \mathbb{R}^{74}$, який включає:

– кількісні метрики: загальна кількість подій автентифікації, число підключень змінних носіїв, кількість файлових операцій, обсяг електронної пошти тощо;

– часові характеристики: час першої та останньої активності, розподіл подій між робочими та неробочими годинами, наявність нічних сесій;

– статистичні показники: середні та сумарні обсяги трафіку, співвідношення зовнішніх та внутрішніх листів;

– специфічні індикатори загроз: копіювання файлів чутливих розширень (.pdf, .doc) на USB-носії, відвідування сайтів пошуку роботи тощо.

Додатково вектор x доповнюється 72-вимірним one-hot поданням статичного контексту (роль, підрозділ), утворюючи повний вектор стану.

$$S \in \mathbb{R}^{146}.$$

Нормалізація даних для навчання GAN

Специфіка навчання GAN висуває суворі вимоги до розподілу даних. Поведінкові дані CERT мають незбалансовані дані: більшість днів характеризується низькою активністю, але трапляються сплески екстремальних значень. Для стабілізації навчання

застосовано двоетапну процедуру.

1. Логарифмічна трансформація. До всіх кількісних ознак v застосовується перетворення, що зменшує асиметрію розподілів і наближає їх до нормального вигляду.

2. Масштабування (Min–Max). Ознаки приводяться до діапазону [-1; 1], що відповідає області значень функції активації tanh у генераторі:

$$\tilde{v} = 2 \cdot \frac{v' - \min(v')}{\max(v') - \min(v')} - 1$$

Це запобігає насиченню градієнтів та забезпечує стабільну збіжність під час навчання мережі.

Генератор CTGAN навчається моделювати умовний розподіл табличних ознак позитивного класу $p(x \mid y = 1, c)$, де $y = 1$ відповідає інсайдерській атаці, а c позначає контекстні змінні (роль, підрозділ тощо). Врахування контексту дозволяє зберігати зв'язок між роллю користувача та характером його дій.

Комбінація логарифмічної трансформації та масштабування в **[-*1*, *1*]** забезпечує для більшості ознак компактний, збалансований діапазон значень без домінування поодиноких екстремальних спостережень. У такому просторі GAN отримує більш керований вхід: градієнти не затухають через надто малі або надто великі аргументи активаційних функцій, а геометрія даних стає більш придатною для відтворення генератором. У результаті процес навчання набуває більш стабільного характеру, а модель краще відтворює реалістичні профілі типової поведінки, на тлі яких аномалії інсайдерської активності проявляються більш контрастно.

Атрибути, що описують роль користувача, його підрозділ та інші дискретні характеристики, кодуються у векторній формі за допомогою методу one-hot кодування. Отриманий вектор умов *y* конкатенується з шумовим вектором *z* на вході генератора, а також із вектором реальних або згенерованих ознак *x* на вході дискримінатора. Оскільки ці дані є статичними умовами і не потребують генерації (зворотного поширення помилки через операцію вибору категорії), використання специфічних технік диференційованої дискретизації (на кшталт Gumbel-Softmax) не є необхідним. Використання one-hot кодування забезпечує чітке розділення категоріальних ознак у просторі умов, дозволяючи нейромережі ефективно вивчати залежності між організаційним контекстом та патернами поведінки.

Для кожного об'єкту "користувач-день" формується композитний вектор опису стану $S \in \mathbb{R}^{146}$, який об'єднує динамічні поведінкові ознаки та статичний контекст. Таким чином, добова активність кожного співробітника у всіх розглянутих каналах спостереження подається у вигляді єдиного 146-вимірного вектору $S = [x, y]$, де $x \in \mathbb{R}^{74}$ відповідає нормованим континуальним поведінковим ознакам, а $y \in \{0,1\}^{72}$ – one-hot-поданню контекстних атрибутів. Таке узгоджене векторне представлення слугує базою для подальшого навчання GAN-моделі.

Після визначення схеми подання даних ключовим питанням стає стратегія формування вибірок для експерименту, яка передбачає спеціальний підхід до поділу на навчальну, валідаційну та тестову множини. Навчальна вибірка (training set) формується таким чином, щоб містити виключно "чисті", нормальні (benign) дні активності користувачів. Для цього на основі доступної розмітки усі події, пов'язані з інсайдерською активністю, ідентифікуються та вилучаються з тренувального набору. Такий підхід дає змогу GAN наближено відтворити латентний розподіл нормальної поведінки P_{data}, не

додаючи до нього аномальні приклади, які могли б викривити оцінку та знизити чутливість до справжніх загроз.

Окремо виділяється валідаційна вибірка, що складається з частини нормальних даних, не використаних безпосередньо під час навчання. Вона слугує для налаштування гіперпараметрів моделі (розмір латентного простору, параметри регуляризації, швидкість навчання тощо), вибору оптимальної кількості епох та реалізації механізму ранньої зупинки. Оскільки валідаційний набір формально "невідомий" моделі, якість реконструкції та значення цільових критеріїв на ньому дають більш об'єктивну оцінку ступеня узгодження генератора з розподілом нормальної поведінки, ніж показники, отримано на тренувальних даних.

Тестова вибірка містить суміш нормальних днів активності (які не з'являлися ні в навчальній, ні у валідаційній вибірках) та всіх доступних інсайдерських інцидентів. Саме на цьому наборі проводиться фінальне оцінювання ефективності детектора, для кожного об'єкта типу "користувач-день" обчислюється міра аномальності (наприклад, на основі помилки реконструкції або оцінки дискримінатора), після чого будуються метрики якості класифікації нормальних та інсайдерських днів (ROC-криві, AUC, показники чутливості та специфічності тощо). Таким чином, тестовий набір відображає реалістичну для експлуатації ситуацію, у якій домінують нормальні спостереження, а інсайдерські епізоди становлять малу, але критично важливу частку.

Описана стратегія формування вибірок забезпечує чистоту експерименту: модель навчається виключно тому, що вважається "нормою", і надалі інтерпретує загрози як статистичні відхилення від вивченого розподілу, не покладаючись на попередньо відомі сигнатури атак. Це повністю відповідає концепції виявлення аномалій.

Таблиця 1

Специфікація ознак для моделі cGAN

Група	К-ть	Опис	Тип
Вектор поведінки x (генерується моделлю)			
Logon	4	Час входу/виходу, тривалість сесії, нічна робота	Log + MinMax
Device	8	USB-активність (робочий/неробочий час)	Log + MinMax
File	32	Файлові операції (.doc, .pdf, .xls, .zip)	Log + MinMax
HTTP	18	Web-категорії, обсяги трафіку	Log + MinMax
E-mail	12	Зовнішні листи, вкладення	Log + MinMax
Підсумок x	74	Континуальні динамічні ознаки	
Вектор умов y (подається на вхід)			
Контекст	72	Роль, Підрозділ, Проєкт, тощо	Категоріальні
Всього	146	Вхідний простір дискримінатора $[x, y]$	

Представлена методика підготовки даних слугує фундаментом для побудови генеративної моделі. Її ключовими елементами є формування інформативного

векторного подання "користувач–день", урахування організаційного та психометричного контексту, а також коректна математична нормалізація ознак із урахуванням специфіки навчання GAN. Сукупність цих кроків дозволяє суттєво зменшити негативний вплив гетерогенності джерел, різнорідних масштабів вимірювання та наявного в даних шуму, перетворюючи журнали подій на узгоджений простір ознак зі стабільними статистичними властивостями, сприятливими для моделювання. На такій основі GAN отримують змогу ефективно вивчати латентну структуру нормальної поведінки користувачів і, відповідно, більш надійно виявляти приховані інсайдерські загрози як відхилення від цього розподілу.

РЕЗУЛЬТАТИ ДОСЛІДЖЕННЯ

На основі датасету CERT r4.2 було сформовано три варіанти тренувальних наборів:

– базовий набір: зберігає природний дисбаланс класів (орієнтовно 1:1000);

– SMOTE-набір: балансування міноритарного класу до співвідношення 1:1 за допомогою SMOTE;

– GAN-набір: балансування до 1:1 за допомогою синтетичних прикладів, згенерованих CTGAN [12].

Тестова вибірка становила 20 % від початкових даних, містила лише реальні події та зберігала вихідний дисбаланс, що забезпечувало чистоту експерименту та відсутність синтетичних зразків у тестувальних даних.

Як класифікатори використовувалися ансамблеві алгоритми:

– Random Forest: налаштовано на велику глибину дерев і кількість базових класифікаторів n = 500 для побудови складної розділяючої поверхні;

– Isolation Forest: алгоритм навчання без учителя, у якому параметр contamination адаптувався до пропорції класів у навчальній вибірці.

Результати тестування було узагальнено в таблицю 2 й підтвердили критичну важливість балансування класів. Навчання Random Forest на незбалансованих даних продемонструвало низьку повноту (Recall = 0.344), тобто більшість атак залишаються невиявленими.

Таблиця 2

Порівняння метрик якості для різних підходів до балансування

Модель	Аугментація даних	Precision	Recall	F1-Score
Random Forest	відсутня	0.624	0.344	0.443
Random Forest	SMOTE	0.895	0.860	0.877
Random Forest	GAN	0.701	0.644	0.671
Isolation Forest	відсутня	0.566	0.624	0.593
Isolation Forest	SMOTE	0.788	0.819	0.803
Isolation Forest	GAN	0.596	0.733	0.657

Використання GAN дозволило суттєво покращити результати порівняно з базовим варіантом: F1-міра для Random Forest зросла з 0.443 до 0.671. Це підтверджує, що синтетичні дані допомагають моделі сформувати кращі правила детектування інсайдерських атак.

Водночас у цьому конкретному експерименті GAN-аугментація поступилася методу SMOTE (F1 = 0.877). Це можна пояснити геометричними властивостями синтезованих даних. SMOTE заповнює простір між найближчими сусідами, створюючи більш щільні кластери, що спрощує завдання класифікатора. Натомість результати GAN можуть страждати від фрагментації або часткового колапсу мод, коли генеруються зразки переважно навколо найбільш типових атак, залишаючи периферійні сценарії без достатнього покриття. Тим не менш, GAN забезпечує логічну валідність згенерованих логів, чого не гарантує SMOTE, який не враховує приховану структуру послідовностей подій.

ВИСНОВКИ ТА ПЕРСПЕКТИВИ ПОДАЛЬШИХ ДОСЛІДЖЕНЬ

У роботі досліджено метод підвищення ефективності детектування інсайдерських загроз шляхом аугментації даних за допомогою генеративно-змагальних мереж. Розроблено повний конвеєр підготовки даних, що включає агрегацію логів до рівня "користувач–день", інженерію поведінкових ознак, логарифмічну нормалізацію та масштабування до діапазону [-1; 1] з урахуванням контексту ролі користувача й підрозділу.

Експериментальні результати на датасеті CERT Insider Threat Dataset r4.2 показали, що використання генеративної моделі CTGAN для збагачення навчальної вибірки дозволяє суттєво підвищити якість виявлення інсайдерів порівняно з навчанням на незбалансованих даних (ріст F1-міри для Random Forest з 0.443 до 0.671).

Попри те, що на табличних даних CERT метод SMOTE показав вищу ефективність за метрикою F1 завдяки щільній інтерполяції між існуючими зразками, генеративні моделі мають важливу перевагу – здатність створювати складні, логічно узгоджені сценарії атак, які неможливо отримати простими геометричними методами. Подальший розвиток методу полягає у вдосконаленні архітектур генераторів для кращого охоплення рідкісних мод розподілу, поєднанні GAN-аугментації з методами навчання з підкріпленням для моделювання послідовностей подій, а також інтеграції з методами пояснюваного штучного інтелекту з метою підвищення прозорості прийняття рішень у системах виявлення інсайдерських загроз.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Greitzer, F. L., & Hohimer, R. E. (2011). Modeling Human Behavior to Anticipate Insider Attacks. *Journal of Strategic Security*, *4*(2), 25–48. http://dx.doi.org/10.5038/1944-0472.4.2.2.
2. 2018 Cost of Insider Threats: Global. Research Report / Ponemon Institute LLC. // Traverse City, MI, 2018. Режим доступу: https://www.insiderthreatdefense.us/pdf/Ponemon%20Institute%202018%20Report%20-%20The%20True%20Cost%20Of%20Insider%20Threats%20Revealed.pdf.
3. Homoliak, I., Toffalini, F., Guarnizo, J., Elovici, Y., & Ochoa, M. (2019). Insight Into Insiders and IT. *ACM Computing Surveys*, *52*(2), 1–40. https://doi.org/10.1145/3303771.
4. ВАЙС, Т., ОНИЩАК, Н., ПОЛОВКО, І., & ШАРКАДІ, М. (2024). ВИКОРИСТАННЯ ГЕНЕРАТИВНОГО ШТУЧНОГО ІНТЕЛЕКТУ ДЛЯ АНАЛІЗУ ДАНИХ. *MEASURING AND*

COMPUTING DEVICES IN TECHNOLOGICAL PROCESSES, (2), 389–394. https://doi.org/10.31891/2219-9365-2024-78-45.

5. Yuan, S., & Wu, X. (2021). Deep learning for insider threat detection: Review, challenges and opportunities. *Computers & Security*, *104*, 102221. https://doi.org/10.1016/j.cose.2021.102221.
6. Dunmore, A., Jang-Jaccard, J., Sabrina, F., & Kwak, J. (2023). A Comprehensive Survey of Generative Adversarial Networks (GANs) in Cybersecurity Intrusion Detection. *IEEE Access*, 1. https://doi.org/10.1109/access.2023.3296707.
7. Preston M. (2022). Insider Threat Detection Data Augmentation Using WCGAN-GP: Master's Thesis // Preston Mack. Halifax, Nova Scotia, Canada: Dalhousie University. http://hdl.handle.net/10222/81531.
8. Gayathri R. G., Sajjanhar, A., & Xiang, Y. (2024). Hybrid deep learning model using SPCAGAN augmentation for insider threat analysis. *Expert Systems with Applications*, 123533. https://doi.org/10.1016/j.eswa.2024.123533.
9. Donahue J., Krähenbühl P. & Darrell T. (2017). Adversarial Feature Learning. Proceedings of the International Conference on Learning Representations (ICLR). https://doi.org/10.48550/arXiv.1605.09782.
10. Chen, Y., Chen, W., Chandra Pal, S., Saha, A., Chowdhuri, I., Adeli, B., Janizadeh, S., Dineva, A. A., Wang, X., & Mosavi, A. (2021). Evaluation efficiency of hybrid deep learning algorithms with neural network decision tree and boosting methods for predicting groundwater potential. *Geocarto International*, 1–21. https://doi.org/10.1080/10106049.2021.1920635.
11. Korchenko, O., Korchenko, A., Zybin, S., & Davydenko, K. (2025). An approach for classifying sociotechnical attacks. *Radioelectronic and Computer Systems, 2025*(2), 230-252. https://doi.org/10.32620/reks.2025.2.15.
12. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. https://doi.org/10.1613/jair.953.
13. Xu L., Skoularidou M., Cuesta-Infante A., Veeramachaneni K. (2019). Modeling Tabular Data Using Conditional GAN. Advances in Neural Information Processing Systems, 7, 7335–7345. https://doi.org/10.48550/arXiv.1907.00503.
14. Liberti, G. (2009). Improved Strategies for Branching on General Disjunctions. *Zootaxa*, *2318*, 339–385. https://doi.org/10.1184/R1/12841247.v1.

КІБЕРБЕЗПЕКА: освіта, наука, техніка
CYBERSECURITY: EDUCATION, SCIENCE, TECHNIQUE
№ 3 (31), 2025
ISSN 2663 - 4023

Vitalii O. Verbynenko
Student of the Department of Cybersecurity Systems and Technologies
State University of Information and Communication Technologies, Kyiv, Ukraine
ORCID: 0009-0004-2134-1987
vv.q@ukr.net

Serhii V. Zybin
Dr. Sci. (Engin.), professor,
professor of the Department of Cybersecurity Systems and Technologies
State University of Information and Communication Technologies, Kyiv, Ukraine
professor of the Department of Applied Information Systems
Taras Shevchenko National University of Kyiv, Kyiv, Ukraine
ORCID: 0000-0002-2670-2823
zysv@ukr.net

METHOD FOR INCREASING THE DETECTION EFFICIENCY OF INSIDER THREATS USING GAN AUGMENTATION

Abstract. In modern corporate information systems, a significant proportion of information security incidents are insider threats. This creates new requirements for security event monitoring and analysis systems. Unlike external attacks, insider activity is disguised as the usual work of legitimate users, and therefore is difficult to describe using classic signature or perimeter protection mechanisms. An additional complexity is the extreme imbalance of classes in event logs. The number of records of typical daily activity is thousands of times higher than the number of recorded incidents. This leads to degradation of the quality of standard machine learning algorithms. The article develops an approach to increasing the efficiency of detecting insider threats by augmenting data using generative adversarial networks, in particular the Conditional Tabular GAN (CTGAN) architecture. A process for preparing behavioral logs is proposed. This process involves the aggregation of multi-channel events to the "user-day" level, construction of a vector of dynamic behavioral features and static context, logarithmic normalization of features with "heavy tails" and scaling to the range [–1; 1]. This ensures stable training of the generative model. CTGAN is configured to simulate the conditional distribution of tabular data of the minority class (insider attacks) taking into account the context of the user's role and department. For each continuous feature, specialized normalization is applied, which allows for the correct reproduction of multimodal distributions, and for discrete variables, the Gumbel-Softmax technique is used, which makes it possible to learn using the backpropagation method of the error. The proposed method is promising for integration into SIEM/UEBA class systems and further combination with methods of explanatory artificial intelligence.

Keywords: insider threats; generative adversarial networks; data augmentation; anomaly detection; information security.

REFERENCES (TRANSLATED AND TRANSLITERATED)

1. Greitzer, F. L., & Hohimer, R. E. (2011). Modeling Human Behavior to Anticipate Insider Attacks. *Journal of Strategic Security*, *4*(2), 25–48. http://dx.doi.org/10.5038/1944-0472.4.2.2.
2. 2018 Cost of Insider Threats: Global. Research Report / Ponemon Institute LLC // Traverse City, MI, 2018. Режим доступу: https://www.insiderthreatdefense.us/pdf/Ponemon%20Institute%202018%20Report%20-%20The%20True%20Cost%20Of%20Insider%20Threats%20Revealed.pdf.
3. Homoliak, I., Toffalini, F., Guarnizo, J., Elovici, Y., & Ochoa, M. (2019). Insight Into Insiders and IT. *ACM Computing Surveys*, *52*(2), 1–40. https://doi.org/10.1145/3303771.
4. VAIS T., ONYSHCHAK N., POLOVKO I. & SHARKADI M. (2024). USING GENERATIVE ARTIFICIAL INTELLIGENCE FORDATA ANALYSIS. *MEASURING AND COMPUTING DEVICES IN TECHNOLOGICAL PROCESSES*, (2), 389–394. https://doi.org/10.31891/2219-9365-2024-78-45.

5. Yuan, S., & Wu, X. (2021). Deep learning for insider threat detection: Review, challenges and opportunities. *Computers & Security*, *104*, 102221. https://doi.org/10.1016/j.cose.2021.102221.
6. Dunmore, A., Jang-Jaccard, J., Sabrina, F., & Kwak, J. (2023). A Comprehensive Survey of Generative Adversarial Networks (GANs) in Cybersecurity Intrusion Detection. *IEEE Access*, 1. https://doi.org/10.1109/access.2023.3296707.
7. Preston M. (2022). Insider Threat Detection Data Augmentation Using WCGAN-GP: Master's Thesis // Preston Mack. Halifax, Nova Scotia, Canada: Dalhousie University. http://hdl.handle.net/10222/81531.
8. Gayathri R. G., Sajjanhar, A., & Xiang, Y. (2024). Hybrid deep learning model using SPCAGAN augmentation for insider threat analysis. *Expert Systems with Applications*, 123533. https://doi.org/10.1016/j.eswa.2024.123533.
9. Donahue J., Krähenbühl P. & Darrell T. (2017). Adversarial Feature Learning. Proceedings of the International Conference on Learning Representations (ICLR). https://doi.org/10.48550/arXiv.1605.09782.
10. Chen, Y., Chen, W., Chandra Pal, S., Saha, A., Chowdhuri, I., Adeli, B., Janizadeh, S., Dineva, A. A., Wang, X., & Mosavi, A. (2021). Evaluation efficiency of hybrid deep learning algorithms with neural network decision tree and boosting methods for predicting groundwater potential. *Geocarto International*, 1–21. https://doi.org/10.1080/10106049.2021.1920635.
11. Korchenko, O., Korchenko, A., Zybin, S., & Davydenko, K. (2025). An approach for classifying sociotechnical attacks. *Radioelectronic and Computer Systems, 2025*(2), 230-252. https://doi.org/10.32620/reks.2025.2.15.
12. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. https://doi.org/10.1613/jair.953.
13. Xu L., Skoularidou M., Cuesta-Infante A., Veeramachaneni K. (2019). Modeling Tabular Data Using Conditional GAN. Advances in Neural Information Processing Systems, 7, 7335–7345. https://doi.org/10.48550/arXiv.1907.00503.
14. Liberti, G. (2009). Improved Strategies for Branching on General Disjunctions. *Zootaxa*, *2318*, 339–385. https://doi.org/10.1184/R1/12841247.v1.