УДК 004.891

д.т.н., проф., **Лєнков С.В.** (ВІКНУ)
к.т.н., доц., **Джулій В.М.** (ХмНУ)
к.т.н., доц., **Муляр І.В.** (ХмНУ)
к.т.н., **Кубявка М.Б.**(ВІКНУ)

DOI: https://doi.org/10.17721/2519-481X/2023/81-09

ФУНКЦІОНАЛЬНА МОДЕЛЬ КЛАСИФІКАЦІЇ ПСЕВДОВИПАДКОВИХ ПОСЛІДОВНОСТЕЙ ЗАШИФРОВАНИХ ТА СТИСЛИХ ДАНИХ ЗАПОБІГАННЮ ВИТОКУ КОНФІДЕНЦІЙНОЇ ІНФОРМАЦІЇ

Розглянута задача побудови формалізованої моделі інсайдера, яка може застосовуватись як у комерційних так і державних компаніях. Показано, що загрози безпеки даних характеризуються набором векторних показників, якісних та кількісних, для їх формалізації необхідне застосування теорії нечітких множин та дискретної математики. Показано неможливість застосування експертних традиційних методів оцінок для визначення більшості розглянутих показників.

Для мінімізації ризику витоку конфіденційної інформації пропонується формувати групи співробітників та розраховувати ризик витоку конфіденційних даних для кожної з них.

Розробка моделі псевдовипадкових послідовностей дозволить оцінити ступінь впливу статистичних ознак, що витягуються з псевдовипадкових послідовностей і використовуються в процесі формування класифікатора, на точність проведення процедури класифікації. Отримані кількісні значення ознак дозволять оптимізувати кількість параметрів за умови дотримання необхідної точності, оцінити складність виконання процедури видалення ознак. На основі отриманих результатів моделювання, виявлених особливостей класифікатора необхідно обґрунтувати вибір математичного апарату, що в подальшому дозволить перейти до практичної реалізації алгоритму класифікації послідовностей, сформованих алгоритмами стиснення та шифрування даних.

Проведений аналіз досліджень у даній предметній області дозволив виявити практичну проблему наявних механізмів захисту: низька точність виявлення зашифрованої інформації, через їх схожість з типовими високоентропійними послідовностями, використання службової інформації притаманної процесу передачі, зберігання конфіденційної інформації. Таким чином задача класифікації зашифрованих та стислих даних є актуальною.

Для вирішення поставленої задачі необхідно: провести аналіз особливостей функціонування перспективних засобів запобігання та виявлення витоку конфіденційних даних, виявити обмеження, пов'язані з виявленням стислої та зашифрованої інформації, обґрунтувати вибір відповідного ознакового простору для моделювання, сформованих алгоритмами стиснення та шифрування інформації, псевдовипадкових послідовностей; розробити модель, сформованих алгоритмами стиснення та шифрування даних, псевдовипадкових послідовностей, що відрізняється від відомих, врахуванням їх статистичних характеристик.

Представлена модель псевдовипадкових послідовностей, відрізняється від аналогів з врахуванням розподілу байт та з врахуванням частот бітових підпослідовностей довжини 9 біт. Для оцінки адекватності запропонованої моделі проведені експерименти щодо визначення точності класифікації псевдовипадкових послідовностей алгоритмами машинного навчання.

Ключові слова: псевдовипадкові послідовності, функціональна модель, інформаційна безпека, точність класифікації, зашифровані, стислі дані.

Вступ. Інформаційні технології розвиваються дуже стрімко, зростає доступність освіти у сфері комп'ютерних наук та високих технологій. На сьогодні отримати доступ до інформації, що дозволяє подолати механізми захисту даних не важко. Людство стикається з інформаційними системами повсюдно: вдома, на роботі, записуючись на прийом до лікаря та отримуючи державні послуги, велика частка персоналу має доступ до даних клієнтів, захищених інформаційних ресурсів, конфіденційної інформації компанії [1 - 3].

Незважаючи на удосконалення механізмів захисту від кіберзагроз, розвиток засобів захисту конфіденційної інформації, зростає кількість витоків конфіденційної інформації. Однією з головних причин зростаючої кількості витоків конфіденційної інформації -

наявність внутрішнього порушника, здатного дотримуватись встановлених правил та заходів роботи з даними, здійснювати передачу конфіденційної інформації за контрольований інформаційний периметр компанії [4, 5, 9, 13].

Забезпечення інформаційної безпеки конфіденційних даних та призупинення дій внутрішнього порушника здійснюється за допомогою організаційних заходів. Виконується тестування, перевірка фактів їхньої біографії, відбір кандидатів, протягом проміжку часу можуть змінитися багато факторів, один з яких - лояльність співробітника [5 - 8, 22, 24].

Проведений аналіз інцидентів інформаційної безпеки, аналітичними центрами компаній SafeNet свідчить про те, що у випадках витоку конфіденційних даних більш ніж 52% винуватцями виявлялися внутрішні порушники [5, 19, 22, 27].

На теперішній час захист від витоків даних реалізується засобами запобігання та виявлення витоку конфіденційної інформації. Основними механізмами захисту від витоків даних, є методи, засновані на пошуку регулярних виразів, сигнатур, цифрових зліпків, виявлення аномалій, застосування алгоритмів машинного навчання [8, 11, 12, 16, 17].

Доступність освіти у сфері високих технологій та розвиток інформаційних технологій визначають широке застосування систем обробки, зберігання, передачі даних та, як наслідок, загрози інформаційної безпеки. У сучасній організації бізнес процеси неможливі без застосування корпоративних мереж передачі даних та перспективних інформаційних систем. З кожним роком збільшуються обсяги інформації, що обробляються, впроваджуються нові інформаційно - пошукові системи, у тому числі системи обробки та збереження конфіденційних даних різного рівня доступу. Якщо механізми захисту даних від зовнішніх загроз досягли відповідних гарантованих рівнів, то способи та методи протидії інсайдеру слабо розвинені, в більшості документів, що регламентують політику безпеки кофіденційним даним компанії, містяться постулати про відсутність інсайдера, що тягне до зростання ймовірності порушення інформаційної безпеки даних, що захищаються [6, 9, 13, 17, 18, 26].

Відповідно до звіту міжнародного експертно-аналітичного центру компаній Group-IB частка інсайдерів, як джерел зареєстрованих випадків в організаціях витоку конфіденційної інформації, за період із січня по червень 2022р. склала понад 80%. У 78% зареєстрованих випадках витоку даних було організовано навмисне [13, 16, 26].

Типовими внутрішніми порушниками є співробітники, які займають технічну позицію - не привілейовані технічні користувачі. Об'єктом атаки є конфіденційні дані організації, такі як програмне забезпечення, фізичне обладнання, бізнес-плани, особливості виробничих процесів, бухгалтерські звіти, бази даних різних рівнів та інші дані, які можуть мати деяку цінність для внутрішнього порушника особисто, або для отримання ділових переваг. Активна діяльність інсайдера, в більшості, триває від одного до чотирьох місяців. Якщо планується звільнення внутрішнього порушника, то в даний період входять наступні події: прийняття рішення про звільнення; період злочинної активності; замітання слідів, щоб мінімізувати ризик виявлення [5, 6, 9 - 11].

Розглянута задача побудови формалізованої моделі інсайдера, яка може застосовуватись як у комерційних так і державних компаніях. Показано, що загрози безпеки даних характеризуються набором векторних показників, якісних та кількісних, для їх формалізації необхідне застосування теорії нечітких множин та дискретної математики. Побудовано формалізовану модель інсайдера, із застосуванням рейтингового методу, засновану на багатокритеріальному ранжируванні. На основі лінгвістичного підходу проведено формалізацію нечіткої інформації з переходом до кількісної єдиної шкали. Також в роботі розглянуто приклад визначення рівня загрози внутрішнього порушника із побудовою семантичних моделей для групи співробітників. Показано неможливість застосування експертних традиційних методів оцінок для визначення більшості розглянутих показників. Проведено аналіз байєсовського підходу вирішення задачі, доведено, при цьому, необхідність проведення аналізу великої кількості статистичних даних. Запропоновано використовувати модель Бьюкенена і Шортліфа, яка дозволяє навести результати на основі використання неповних відомостей про об'єкт, що аналізується [9, 13, 17, 18, 22].

Актуальність інсайдера визначається рейтинговою оцінкою, його становищем в рейтингу. Багатокритеріальне ранжування передбачає групове ранжування (класифікацію, кластеризацію) - віднесення співробітників на основі лінійного ранжування до упорядкованих груп. Головна перевага рейтингового підходу – комплексний характер до оцінки рівня інсайдерської безпеки. Рейтинговий метод має низку істотних недоліків: неможливість застосування однакових арифметичних операцій для значень показників моделі внутрішнього порушника, що вимірюються у якісних та кількісних шкалах; у зв'язку з тим, що модель внутрішнього порушника містить велику кількість показників, які можуть мати кореляційні зв'язки між собою, що впливають на рівень інсайдерської безпеки, виникають, в даній ситуації, труднощі в комплексному підході оцінки рівня інсайдерських атак та загроз по окремих співробітниках; відсутня формалізована процедура визначення значень кількісних та якісних показників; використана в неформалізованій моделі інсайдера природна мова зрозуміла аналітику, добре передає семантику предметної області, але не дозволяє однозначно і точно описати взаємозв'язки сутностностей, представлені в моделі внутрішнього порушника [14, 21 - 23].

У зарубіжних дослідженнях наголошується на необхідності прийняття відповідних заходів щодо протидії інсайдерам. Згідно зі статистикою Національного центру безпеки Південнокорейської республіки близько 75% витоків конфіденційної інформації відбувається з вини поточних або колишніх співробітників компанії. Більшість витоків конфіденційних даних відбувається через недосконаліть засобів з їх виявлення і запровадження недостатніх заходів щодо припинення витоків інформації. Більшість робіт із забезпечення інформаційної безпеки конфіденційних даних пов'язані із захистом від проведеннія зовнішніх атак, що підтверджує актуальність проведеного аналізу досліджень [6, 7, 18, 19, 22].

Основними джерелами загроз та атак для корпоративних мереж є: технічні, що відносяться до особливостей обслуговування, функціонування, створення програмно-апаратних, апаратних, програмних засобів; суб'єктивні, викликані відповідними діями співробітників компанії. У наведених групах є підклас джерел, що відноситься до інсайдерів. Відзначається також наявність загроз та атак промислового шпигунства, що реалізується шкідливим програмним забезпеченням чи внутрішнім порушником, також різних botnet мереж. Основним засобом поширення та зараження шкідливого програмного забезпечення є botnet мережі. Відзначається можливість передачі інсайдерами захищених даних з контрольованого периметра компанії з використанням сервісів електронної пошти [7 - 11, 17].

Для мінімізації ризику витоку конфіденційної інформації пропонується формувати групи співробітників та розраховувати ризик витоку конфіденційних даних для кожної з них. Запропонований підхід передбачає використання data leakage prevention (DLP) та security information and event managment (SIEM) систем. Причиною витоку даних можуть бути політичні, індивідуальні, фінансові мотиви працівників компанії [5, 6, 13, 18, 19].

Аналіз останніх досліджень та постановка задачі. Аналіз досліджень мережевої активності корпоративної мережі є ключовим компонентом запобігання та раннього виявлення загроз та атак безпеки конфіденційним даним, що виходять від інсайдерів. Логування подій безпеки та функціонування інформаційної системи можуть використовуватись у реальному часі для проведення аналізу, проте записи необхідно відфільтровувати, оскільки не всі дозволяють виявити загрозу, атаку безпеки даних. Для розробки моделі псевдовипадкових чисел необхідно розглянути результати аналізу досліджень предметної області, визначити які ознаки найчастіше використовувані, що описують стислі та зашифровані послідовності. Розробка моделі псевдовипадкових послідовностей дозволить оцінити ступінь впливу статистичних ознак, що витягуються з псевдовипадкових послідовностей і використовуються, надалі в процесі формування класифікатора, на точність проведення процедури класифікації. Отримані кількісні значення ознак дозволять оптимізувати кількість параметрів за умови дотримання необхідної точності, оцінити складність виконання процедури видалення ознак. На основі отриманих результатів

моделювання, виявлених особливостей класифікатора необхідно обґрунтувати вибір математичного апарату, що в подальшому дозволить перейти до практичної реалізації алгоритму класифікації псевдовипадкових послідовностей, сформованих алгоритмами стиснення та шифрування даних [5, 15 - 19].

Проведений аналіз відкритого та зашифрованого трафіку на основі підрахунку ентропії окремих слів довжиною 2..64 біт, потоку даних, стандартного відхилення та середнього значення зазначених величин. Для проведення експериментів використані пакети довжиною понад 20 байт: відкритих та зашифрованих. Найкращі результати досягнуті при використання алгоритму C5.0, точність класифікації - 0,978, використовувалися пакети певних додатків і протоколів: skype, https, smtp, dtls, ssl, ldap, http, youtube, decphone, netbios, dns, виявлено у них службові специфічні ознаки, які дозволяють класифікувати з високою точністю трафік.

Широке впровадження функцій шифрування інформації при передачі даних призводить до того, що використовувані методи класифікації зашифрованого трафіку не справляються з задачами з високою точністю, наприклад, при класифікації потоку даних, що мають схожі ознаки, цифрові зліпки. Пропонується метод класифікації зашифрованого потоку даних на основі застосування ланцюгів Маркова та атрибутів. Для збільшення точності класифікатора застосовуються наступні ознаки: довжина перших даних додатків, довжина сертифіката у SSL/TLS сесіях. Запропоноване рішення дозволило досягти точності класифікації зашифрованого потоку даних 0,907. Недоліками даного рішення, у випадках змін додатків, неможливість правильно їх класифікувати, оскільки зміняться значення біграм, додатки, які не брали участь у навчання класифікатора також неможливо класифікувати. Для класифікації зашифрованого потоку даних від 10 різних Інтернет ресурсів, запропонований алгоритм обчислення відстані між класами, що обробляється методом k-найближчих сусідів. При побудові ознакового простору використані статистичні ознаки: службові характеристики пакетів (міжінтервальний час пакетів і довжина), мережеві характеристики трафіка (IP-адреса, номер порту, кількість пакетів, тривалість потоку), дані встановлення TLS з'єднання (довжина публічного ключа, відповідь сервера), характеристики розподілу байт. Середня точність алгоритму склала 0,947, точність алгоритму побудови ймовірного лісу - 0,84, алгоритму побудови дерева рішень - 0,879.

Методи глибокого аналізу трафіка, засновані на сигнатурному пошуку, не здатні виявляти зашифрований потік даних, також їх особливістю є складність у проведенні класифікації стислих і зашифрованих даних, використовується розподіл відстані Хемінга для перших байт, IP-телефонії зашифрованого потоку даних, розподіл є біномним з піком у значенні 4 біта, дозволяє, для зашифрованих даних, зробити висновок про рівномірний розподіл біт. Для незашифрованого потоку даних IP-телефонії розподіл має пікоподібну форму з максимумом у значенні 0 біт. Даний підхід застосовується для перших 100 пакетів, наступні пакети мають подібний розподіл і не можуть застосовуватися для класифікації.

Останнім часом дослідження спрямовані на вирішення задачі класифікації зашифрованого потоку даних, при класифікації трафіку додатків та різних протоколів вдалося досягти високих результатів, застосовуючи, при цьому розміри пакетів, величини міжпакетних інтервалів, службову інформацію, номери портів. Технологія глибокого аналізу трафіка не застосовується для інкапсульованого чи зашифрованого потоку даних і на теперішній час є неефективною. Запропонована модель класифікатора трафіка ґрунтується на прихованих ланцюгах Маркова, використовуючи, в даному випадку, як ознаки, розміри пакетів, величини міжпакетних інтервалів, та корелювані за допомогою моделі Гауссових розподілів. Даний підхід дозволив досягти точності класифікації потоку даних - 0,989.

У роботі [20] досліджували можливість класифікації файлів найбільш популярних форматів: doc, csv, docx, gz, gif, jpg, html, png, pdf, pptx, ppt, ps, swf, rtf, xls, txt, xml, xlsx. Для формування ознакового простору використані підпослідовності розміром 4, 8, 16, 32, 64 байт, на основі послідовностей будувалися словники інформації, що містять 1024, 1296, 1444 елемента, словників допускалося накладення словників. Далі здійснювався підрахунок частот знаходження підпослідовностей для кожного типу потоку даних. Для побудови

класифікатора обраний метод опорних векторів з параметром - 32, метод показав найвищу точність класифікації при використанні даних підпослідовності довжиною 64 байт і розміром словника 1444. Точність класифікації, при цьому, досягла 0.617. Виявлення підроблених (спотворених) форматів файлів є актуальним завданням у комп'ютерній криміналістиці - приховування важливої для розслідування інформації (файлів), частини інформації у файлах-контейнерах може затягнути суттєво час розслідування, як наслідок, час реакції на інцидент.

Для класифікації зображень форматів pdf, png, jpeg, gif із заміною магічних байт на користь комп'ютерної криміналістики та внесеними змінами до розширення файлів, як ознаки використовувався за максимальним значенням нормалізований розподіл байт. Для процедури класифікації використовувалися нейронні мережі, для відбору інформативних ознак застосовувалися генетичні алгоритми. Ознаки обчислювалися в середовищі MatLab, відбір ознак генетичним алгоритмом (100 поколінь, популяція з 256 одиниць, перехресне значення 0,8, ймовірність мутацій 0,034), навчання нейронної мережі (швидкість навчання 0,3, 42 входи, 1 прихований шар з 3 вузлами) виконані у програмному середовищі Weka [21, 23]. Точність класифікації зі змінених зображень формату png - 97,91%, для jpeg і gif – 99,99%, для tiff - 98,31%.

При класифікації спотворених файлів 4 класів jpg, gif, png, pdf які входять у набір даних ImageCLEFsecurity, в якості знакового простору використано розподіл байт. Застосовувалися згорткові нейронні мережі ResNet (точність класифікації 0,9997) та VGG-16 (точність класифікації 0,9993). Визначення типу файлів по заголовку і розширенню, по магічним байтам, які містяться в перших 2 - 46 байтах файлу, є ненадійним методом, оскільки дана інформація може бути легко змінена. Для отримання інформативних ознак, застосували метод, який дозволяє перевести частоти зустрічальності в файлах слів в числову квадратну матрицю. Далі було використано метод TF-IDF для вирівнювання статистики розподілу слів у досліджуємих файлах. Для побудови класифікатора потоку даних використовувалися алгоритми машинного навчання та отримані наступні результати: дерево рішень 95,76%, k-найближчих сусідів 37,58%, випадковий ліс 91,86%, метод опорних векторів 68,7%, лінійна регресія 96,46%, алгоритм XGBoost 97,74%, мультиноміальний байєсський класифікатор 96,72%.

Витік інформації є порушенням безпеки даних - порушенням властивості конфіденційності. Зросла цінність, в сучасному суспільстві не тільки даних, що захищаються державою, також персональні дані, корпоративна інформація, позови за розголошення яких становлять мільйони доларів.

Для запобігання реалізації атак та загроз витоку конфіденційної інформації в корпоративних мережах застосовують засоби запобігання та виявлення витоку даних (DLP–системи), які є елементом інформаційної системи безпеки корпоративних мереж. DLP-системи дозволяють знизити ризик реалізації атак та загрози витоку інформації. Однак деякі моделі інсайдерів, які застосовуються в компаніях, також у державних, не містять вимог і заходів захисту від внутрішніх зловмисників. Наведений факт може бути однією з причин збільшення частки інсайдерів у разі витоку конфіденційної інформації [7 - 10, 14, 18, 19].

Відсутність у корпоративній мережі моделей атак та загроз інформації внутрішнього зловмисника обумовлюється проведенням організаційних заходів: визначення посадових співробітників, відповідальних за забезпечення інформаційної безпеки даних; проведення контролю виконання вимог нормативних документів, які регламентують забезпечення захисту кофіденційних даних; встановлення порядку допуску співробітників для проведення відновлювально - ремонтних робіт програмних та технічних засобів; порядку оновлення антивірусних баз; встановлення порядку резервного копіювання, архівування та відновлення баз даних, що знаходяться на різних мережевих рівнях ієрархії компанії.

Наведених заходів недостатньо, у разі наявності в компанії інсайдера. Виявлення внутрішніх порушників організаційними заходами дуже важко, а технічні заходи можуть сприяти розслідуванню інциденту безпеки інформації, але у разі виявлення та затримання зловмисника.

Одним із можливих способів передачі, за периметр організації, даних дотримання встановлених правил безпеки, передача інформації в стислому або зашифрованому вигляді. На теперішній час існують способи класифікації стислої та зашифрованої інформації, однак вони мають низку недоліків.

Кібератаки, особливо ті, які націлені на інформаційні системи обробки та зберігання конфіденційних даних, стають все більш підготовленими та професійними. Критичні національні інфраструктури стають основними об'єктами кібератак, в них обробляється і зберігається найважливіша інформація, захист якої стає проблемою, як для компаній, так і держав. Атаки на такі критичні інформаційні системи включають проникнення в мережу організації та встановлення шкідливого програмного забезпечення, які можуть розкрити конфіденційну інформацію, змінити поведінку технічного обладнання [13, 22, 24].

Дана проблема загострилася останнім часом з огляду на діяльність інсайдерів. Щоб впоратися з цією тенденцією, розробляються нові механізми та системи, які можуть захистити інформаційні системи обробки даних. Поряд з механізмами безпеки, такими як автентифікація, контроль доступу, системи виявлення вторгнень та системи протидії витокам інформації розгортаються як друга лінія оборони. Системи виявлення вторгнень не можуть протидіяти інсайдерам, оскільки націлені на інші методи та механізми, які використовуються зловмисниками. Засоби запобігання, виявлення витоку даних повинні забезпечувати високу швидкість виявлення та низьку частоту помилкових тривог, не вимагаючи, при цьому, значних обчислювальних потужностей для класифікації інформації [7, 13, 18, 19].

Для вирішення поставленої задачі необхідно: провести аналіз особливостей функціонування перспективних засобів запобігання та виявлення витоку конфіденційних даних, виявити обмеження, пов'язані з виявленням стислої та зашифрованої інформації, обґрунтувати вибір відповідного ознакового простору для моделювання, сформованих алгоритмами стиснення та шифрування інформації, псевдовипадкових послідовностей; розробити модель, сформованих алгоритмами стиснення та шифрування даних, псевдовипадкових послідовностей, що відрізняється від відомих, врахуванням їх статистичних характеристик.

Проведений аналіз об'єкта дослідження та предметної області дозволяє висунути гіпотезу про наявність у стислих та зашифрованих даних статистичних особливостей. В результаті досліджень у разі справедливості висунутого припущення можливо сформувати модель псевдовипадкових послідовностей, сформованих алгоритмами стиснення та шифрування потоку даних і розробити метод захисту конфіденційних даних від витоків інформації на основі розподілу зазначених типів даних.

Аналіз досліджень в області інформаційної безпеки щодо внутрішніх зловмисників дозволяє сформувати модель атак та загроз конфіденційним даним за допомогою організації її витоку інсайдером.

Проведений аналіз досліджень у даній предметній області дозволив виявити практичну проблему наявних механізмів захисту: низька точність виявлення зашифрованої інформації, через їх схожість з типовими високоентропійними послідовностями, використання службової інформації притаманної процесу передачі, зберігання конфіденційної інформації. Таким чином задача класифікації стислих та зашифрованих даних є актуальною.

Функціональна модель класифікації псевдовипадкових послідовностей. Процес формування класифікатора послідовностей представляє сукупність дій з виділення псевдовипадкових послідовностей статистичних ознак, обробку послідовностей, вибору відповідного математичного апарату, пошуку найбільш інформативних ознак, навчання класифікатора, проведення процедури тестування отриманого класифікатора.

На рис. 1 представлена схема процесу витоку даних в стислому/зашифрованому виді за контрольований периметр компанії.

Внутрішній порушник (інсайдер) - привілейований користувач, рядовий співробітник компанії, шкідливе програмне забезпечення, встановлене всередині контрольованого периметра підприємства на робочій станції. Передача даних може здійснюватися

стандартними способами та засобами, так і за допомогою засобів стиснення/шифрування інформації.

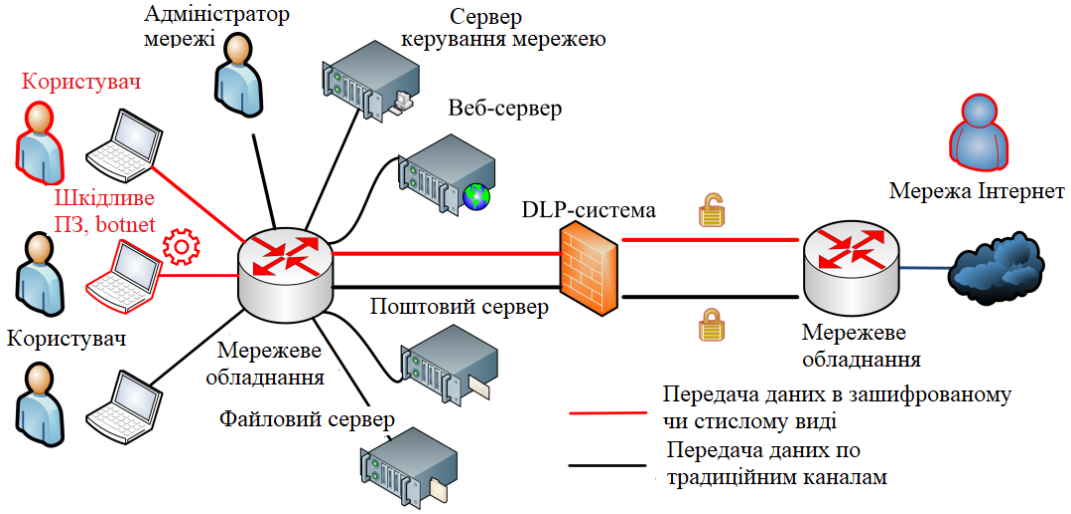

Рисунок 1 - Схема процесу витоку даних, реалізована внутрішнім порушником

У разі використання засобів стиснення/шифрування даних існуючі засоби запобігання та виявлення витокам інформації дозволяють здійснити передачу інформації інсайдеру через наявність практичної проблеми, що полягає в використанні заголовків файлів та низькій точності класифікації даних.

Функціональна модель формування класифікатора представлена на рис. 2. Перед початком процесу формування класифікатора вхідна вибірка потоку даних розбивається на 2 групи підвибірок: тестова, що включає 20% екземплярів вхідної вибірки псевдовипадкової послідовності і навчальна, що складається з 80% екземплярів вхідної вибірки.

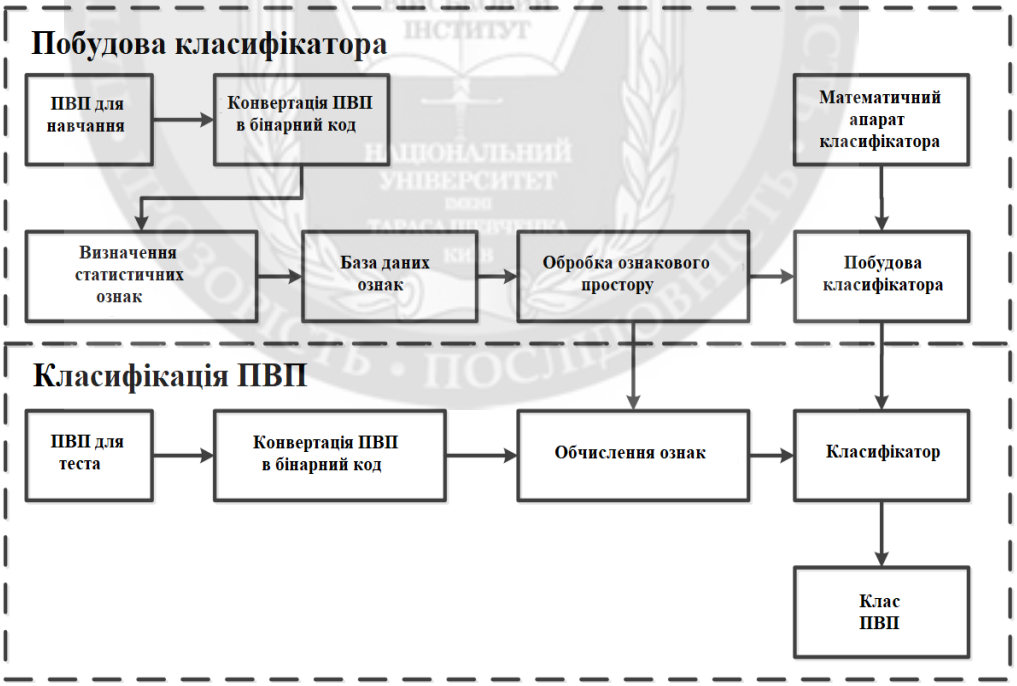

Рисунок 2 - Функціональна модель процесу формування класифікатора

Для розподілу вхідної вибірки псевдовипадкової послідовності застосовується процедура стратифікованого вибору груп, що дозволяє отримати різні набори потоку даних для проведення навчання класифікатора (рис. 3).

Рисунок 3 - Процедура стратифікованого вибору груп

В подальшому необхідно провести конвертацію аналізованих файлів, з відкиданням перших 10 кбайт, у бінарний вигляд для забезпечення незалежності, що розробляється, моделі псевдовипадкових послідовностей, від цифрових сигнатур, що містяться в заголовній частині зашифрованих та стислих файлів. Введення подібної функції дозволяє розглядати файли, як бінарні послідовності.

Для визначення інформаційних ознак, що найбільш повно описують псевдовипадкові послідовності та дозволяють використовувати послідовності для класифікації псевдовипадкові значення, виконується функція визначення ваг послідовностей, яка реалізує обчислення відповідних значень точності проведеної класифікації для кожного класу псевдовипадкових послідовностей та проведення для сформованого класифікатора процедури тестування. Сформований ознаковий простір, що дозволяє, для послідовностей, досягти найбільшої точності класифікації може бути використаний як модель псевдовипадкових послідовностей, згідно з певними метриками.

Обробка отриманих інформаційних ознак, видалення викидів даних і аномальних значень дозволить підвищити точність та якість класифікації псевдовипадкових послідовностей. На наступному кроці необхідно визначити відповідний математичний апарат, що дозволить досягти максимальної якості та точності класифікації псевдовипадкових послідовностей. Алгоритми машинного навчання поділяються на декілька класів, у магістерській роботі розглядаються непараметричні (дерево рішень, випадковий ліс), метричні, що використовують для визначення класу псевдовипадкових послідовностей відповідну метрику відстані (алгоритм *kNN*), також необхідно вибрати метрику точності класифікації псевдовипадкових послідовностей.

Вибір метрики точності класифікації псевдовипадкових послідовностей. Найбільш поширеними та застосовуваними метриками для оцінки результатів експериментів отриманих з машинного навчання є: Precision, Recall, Accuracy, F-міра та AUCROC [10 - 12]. Вибір відповідної метрики залежить від типу розв'язуваної задачі, може фокусувати налаштування класифікатора на набір даних з певними характерними рисами.

Перед вибором метрик визначення точності класифікації псевдовипадкових послідовностей алгоритмами машинного навчання необхідно, покладену в основу цих метрик, розглянути одну з головних концепцій – матрицю помилок. Для випадку використання бінарної класифікації матриця помилок представлена на рис. 4. На рис. 4 наведена графічна інтерпретація метрик Recall (повнота) і Precision (точність).

Як наведено на рис. 4 матриця помилок представляє собою матрицю розміру $N \times N$, де N – кількість класів псевдовипадкових послідовностей, що беруть участь у проведенні

класифікації, у випадку бінарної класифікації $N = 2$. Стовпці матриці представлять множину передбачених класифікатором псевдовипадкових послідовностей. Під час проведення класифікації значення комірок матриці інкрементуються на 1.

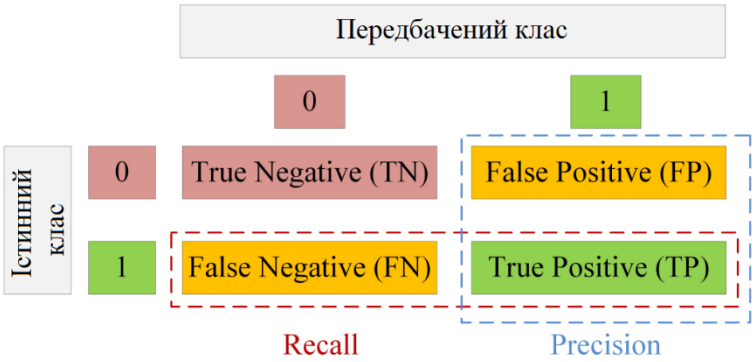

Рисунок 4 - Матриця помилок під час проведення бінарної класифікації псевдовипадкових послідовностей

В залежності від істинності класу псевдовипадкової послідовності, що аналізується, результат оцінки передбаченого класифікатором класу розподіляється на одну з чотирьох груп: вірно віднесених до класу 1 (TP); вірно віднесених до класу 0 (TN); невірно віднесених до класу 1 (FP); невірно віднесених до класу 0 (FN).

Умовний приклад класифікації псевдовипадкових послідовностей представлений на рис. 5. Область вибору класифікатора представлена зафарбованим колом, повні класи псевдовипадкових послідовностей є колонками з виділеними синім і червоним кольором об'єктів. Виходячи з графічної інтерпретації та представленого опису можна зробити висновок про те, що метрика Recall не характеризує здатність класифікатора класифікувати обидва класи псевдовипадкової послідовності, є відображенням здатності знаходити об'єкти певного класу. Метрика Precision дозволяє оцінити, скільки об'єктів одного класу обраних класифікатором дійсно відносяться до заданого. Помилки класифікації поділяють на два типи: False Negative (помилково негативні), False Positive (помилково позитивні). У статистиці перший тип помилок - помилки II-го роду, другий - помилкою I-го роду.

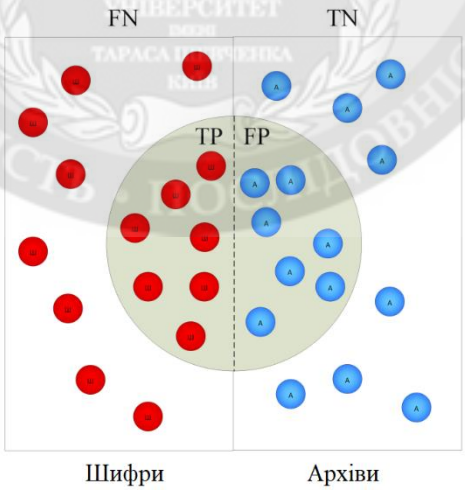

Рисунок 5 - Приклад класифікації псевдовипадкових послідовностей

Метрики точність, повнота, частка правильних відповідей, у формальному виді представлені виразом (1):

$$Precision = \frac{TP}{TP + FP},$$
$$Recall = \frac{TP}{TP + FN}, \quad (1)$$
$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN},$$

де TP - вірно класифіковані класифікатором об'єкти першого класу, TN - класифіковані вірно об'єкти другого класу, FP - класифіковані невірно як об'єкти першого класу об'єкти другого класу, NF - класифіковані невірно як об'єкти другого класу об'єкти першого класу.

Таким чином найбільші значення точності і повноти свідчать про високу здатність класифікатора, одночасно дані метрики максмізувати неможливо. Для оцінки точності класифікатора застосовують метрику F -міра, яка є гармонійним середнім між точністю і повнотою, дозволяє враховувати обидві характеристики (метрики). У разі застосування коефіцієнта β = 1 формула (1) набуває виду F -міри (2). Метрика $g - mean$ (3), враховує правильні відповіді обох класів.

$$F1 = \left(1 + \beta^2\right) \cdot \frac{Precision \cdot Recall}{\left(\beta^2 \cdot Precision + Recall\right)}, \quad (2)$$

$$g_{mean} = \sqrt{\left(\left(\frac{TP}{TP + FN}\right) \cdot \left(\frac{TN}{TN + FP}\right)\right)}, \quad (3)$$

Метрика *Accuracy* (частка правильних відповідей) враховує помилки класифікатора та всі правильні відповіді. Оскільки вибірка даних є збалансованою, дана метрика обрана для проведення експериментів.

Математичний апарат формування класифікатора псевдовипадкових послідовностей. При проведені дослідження в предметній області зустрічаються різні набори даних (відкриті набори даних, згенеровані вибірки даних) для проведення тестування алгоритмів машинного навчання, в обох випадках відсутнє обґрунтування розміру вибірок даних. Для підтвердження статистичної значущості експериментів, необхідно визначити кількість файлів кожного класу, нижню межу розміру вибірки даних, оскільки функції статистичного аналізу для отримання ознак з інформації, потребують часових витрат. Для визначення мінімальної кількості даних у вибірці скористаємося z-критерієм, значення критерія для 99% довірчого інтервалу 2,577 [21]. Оскільки значення статистичних параметрів, дисперсії отримуваних частот підпослідовностей обмеженої довжини N невідомі, скористаємося планованим значенням дисперсії, визначається δ^2, крім того $\partial f = n - 1$, значення n і $t_{\alpha/2\partial f}$ невідомі, значення t-критерія може бути апроксимовано двостороннім z-критерієм. Кількість файлів у кожній групі визначається наступним виразом (4):

$$n = 4 \times \delta^2 \times \left(z_{\alpha/2} / \omega\right)^2 + z_{\alpha/2}^2 / 2 \quad (4)$$

Поправочний коефіцієнт $z_{\alpha/2}^2$ вноситься у разі, що значення критерія z менше значення t-критертерію. Виходячи з проведених експериментів, визначена довжина підпослідовності - 9 біт. Грунтуючись на описі статистичного тесту на підрахунок не перетинаючих шаблонів пакету тестів NIST [21], визначимо середньоквадратичне відхилення і математичне очікування для підпослідовностей довжиною 9 біт (5):

$$\mu = \frac{M - m + 1}{2^m} \quad (5)$$

$$\sigma^2 = M \times \left(\frac{1}{2^m} - \frac{2 \cdot m - 1}{2^{2m}} \right), \quad (6)$$

де M - довжина аналізованої послідовності в бітах, m-довжина підпослідовності в бітах.

Підставляючи отримані значення для довжини послідовності M =700 і m=9 отримаємо значення математичного очікування μ=9600 та середньоквадратичного відхилення σ^2 = 18,1, враховуючи вираз (4) мінімальний розмір кожної групи файлів у вибірці дорівнює 8689 файлів.

Для формування вибірки відібрані текстові файли, що містять осмислений текст українською мовою. Враховуючи специфіку запропонованого підходу та для об'єктивності класифікації, необхідно забезпечити рівність розмірів файлів, що класифікуються, вихідних послідовностей алгоритмів стиснення і шифрування даних, було сформовано дві групи файлів: зашифровані (алгоритми шифрування Camellia, AES, RC4, 3DES), стислі (файли з розширення: ZIP, RAR, GZ, BZ, XZ).

Для побудови класифікатора послідовностей необхідно обґрунтувати вибір використовуваного математичного апарату, який буде використаний для реалізації класифікатора. На теперішній час існує безліч алгоритмів машинного навчання, які демонструють, при вирішенні широкого кола задач, високі результати. В табл. 1 наведено набір файлів проведення експериментів, для вибору математичного апарату класифікатора.

Таблиця 1

Набір файлів проведення експериментів

Алгоритм перетворення	Мітка класу	Кількість файлів	Розмір файла, кБ
AES (CBC)	0	2000	600
3DES (CBC)	0	2000	600
RC4 (CBC)	0	2000	600
RAR	1	2000	600
ZIP	1	2000	600
7Z	1	2000	600
BZ2	1	2000	600

Ретроспективний аналіз дозволяє виділити алгоритми машинного навчання, які застосовуються при класифікації стислих та зашифрованих даних, що найбільш часто зустрічаються в літературі [10].

Застосовувався алгоритм kNN, для класифікації бінарних файлів, точність класифікації відкритих, стислих/зашифрованих даних становила 0,97. Для виявлення зашифрованого трафіку SSH застосовувалися алгоритми адаптивного бустингу та побудови дерева рішень, дозволили досягти точності 0,72. Передана інформація, для обчислення ентропії блоків даних, піддавалася попередньої обробки і передачі класифікатору отриманих ознак для класифікації на основі використання методу опорних векторів, точність дорівнювала 0,79 для стислих/зашифрованих даних.

Для класифікації зашифрованих, бінарних, текстових даних застосовувалися алгоритми побудови дерева рішень та алгоритм опорних векторів, точність становила 0,89. У роботі [14] досліджується можливість класифікації стислих та зашифрованих даних згортковими нейронними мережами алгоритмом і kNN. Найбільшої точності, в даному випадку, досягли використання згорткових нейронних мереж з точністю класифікації 0,68. Досліджується можливість застосування алгоритмів машинного навчання на задачах класифікації не VPN трафіка і VPN, у найбільш пізніх роботах.

Також, розглядаються алгоритми машинного навчання, які показали наступні точності класифікації: випадковий ліс - 0,91; дерево рішень - 0,91; логістична регресія - 0,92; класифікатор на основі згорткових нейронних мереж - 0,97; наївний класифікатор Байєса -

0,35. Застосовувалися метод опорних векторів та алгоритми побудови випадкового лісу, показали невисоку точність при класифікації різних типів файлів: аудіо - 0,63; відео-файли - 0,71; зображення та текстові файли - 0,74.

Таким чином, на теперешній час, можна зробити висновок про найбільш часто використовувані алгоритми машинного навчання: градієнтний бустинг; випадковий ліс; дерево рішень; *kNN*.

У ряді робіт зазначається, що найбільшу точність у задачах класифікації безперервних та категоріальних значень мають алгоритми дерев рішень та побудови випадкового лісу, також у задачах бінарної класифікації. З цієї причини алгоритм *kNN,* дерев рішень, побудови випадкового лісу були обрані для побудови класифікаторів та проведення оцінки їх точності.

Для оцінки адекватності запропонованої моделі проведені експерименти щодо визначення точності класифікації псевдовипадкових послідовностей алгоритмами машинного навчання, отримані результати представлені в табл. 2.

Таблиця 2

Оцінка точності класифікації ПВП алгоритмами машинного навчання при використанні моделі на основі підпослідовностей байт 9 біт

№ п/п	Алгоритм	*Accuracy*
1	Random Forest	0,895
2	Decision Tree	0,892
3	kNN	0,92

Висновки. Представлена модель псевдовипадкових послідовностей, що відрізняється від аналогів з врахуванням розподілу байт та з врахуванням частот бітових підпослідовностей довжини 9 біт.

Проведено аналіз ознакових просторів, які найчастіше використовуються класифікаторами під час класифікації псевдовипадкових послідовностей, обґрунтовано вибір довжини підпослідовностей, розміром в дев’ять біт.

Найбільш часто використовувані алгоритми машинного навчання, в різних дослідженнях, пов'язаних з класифікацією інформації: градієнтний бустинг; випадковий ліс; дерево рішень; *kNN*

Для оцінки адекватності запропонованої моделі проведені експерименти щодо визначення точності класифікації псевдовипадкових послідовностей алгоритмами машинного навчання.

Найбільшу точність у задачах класифікації безперервних та категоріальних значень мають алгоритми дерев рішень та побудови випадкового лісу, також у задачах бінарної класифікації. З цієї причини алгоритм *kNN,* дерев рішень, побудови випадкового лісу були обрані для побудови класифікаторів та проведення оцінки їх точності.

ЛІТЕРАТУРА:

1. Доктрина інформаційної безпеки України, затвердженої Указом Президента України від 25 лютого 2017 року № №47/2017, 15с.

2. Державний стандарт України Захист інформації. Технічний захист інформації. Основні положення. ДСТУ 3396.0-96 [Електронний ресурс]. – Режим доступу : http://www.dsszzi.gov.ua/dsszzi/control/uk/publish/article?art_id=38883&cat_id =38836

3. Бем, М. В. Стандарти захисту персональних даних в соціальній сфері. / М. В.Бем, І. М. Городиський -Львів:, 2018р. - 110 с.

4. Богуш, В.М. Інформаційна безпека держави / В.М. Богуш, О.К. Юдін. – К.: МК-Прес, 2015. – 432 с.

5. Голубєв, О.В. Програмно-технічні засоби захисту даних від комп’ютерних злочинів / О. В. Гошубєв– Запоріжжя : «Павел», 2018. – 145 с.

6. Горбулін, П.В. Проблеми захисту інформаційного простору України / М.М. Баченок, П.В. Горбулін – К.: Інтертехнологія, 2019. – 138 с.

7. Ленков, С.В. Метод прогнозування вразливостей інформаційної безпеки на основі аналізу даних тематичних інтернет-ресурсів / С.В. Лєнков, В.М. Джулій, А.М. Берназ, І.В. Муляр, І.В. Пампуха // Збірник наукових праць Військового інституту Київського національного університету імені Тараса Шевченка. – К.: ВІКНУ, 2023. – Вип. №78. – С. 123-134.

8. Ленков, С.В. Метод протидії поширенню та виявлення шкідливої інформації в соціальних мережах/ С.В. Ленков, В.М. Джулій, Л.В. Солодєєва // Збірник наукових праць Військового інституту Київського національного університету імені Тараса Шевченка. – К.: ВІКНУ, 2022. – Вип. №77. – С. 103-117.

9. Ленков, С.В. Модель безпеки поширення забороненої інформації в інформаційно-телекомунікаційних мережах / С.В. Ленков, В.М. Джулій, В.С. Орленко, О.В. Сєлюков, А.В. Атаманюк // Збірник наукових праць Військового інституту Київського національного університету імені Тараса Шевченка. – К.: ВІКНУ, 2020. – Вип. №68. – С. 53-64.

10. Джулій, В.М. Алгоритми прогнозування вразливостей та загроз інформаційної безпеки на основі тематичних інтернет-ресурсів/ Майор Є., Джулій В., Чешун В., Петляк Н. Міжнародний науково-технічний журнал «Вимірювальна та обчислювальна техніка в технологічних процесах». 2023. Випуск 4. С.49-56.

11. Лєнков, С.В. Інформаційно-аналітична системи прогнозування вразливостей та загроз інформаційної безпеки/ С.В. Лєнков, В.М. Джулій, О.В. Мірошніченко, В.О. Браун, С.І. Прохорський // Збірник наукових праць Військового інституту Київського національного університету імені Тараса Шевченка. – Київ: ВІКНУ, 2023. – Вип. № 79. – С. 114-127

12. Ємельянов, С.Л. Основи інформаційної безпеки. / С.Л. Ємельянов– Одеса: Фенікс, 2019р.– 357 с.

13. Кудінов, В.А. Основи протидії кіберзлочинності. / В. М. Смаглюк, В. Г. Хахановський, В.А. Кудінов. – К. : НАВС, 2016р. – 104 с.

14. Лук'янов, Б. В. Комп'ютерний аналіз даних / Б. В. Лук'янов – К. : Академія, 2017р. – 345 с.

15. Соціальні мережі – реальні загрози віртуального світу. [Електронний ресурс]. – Режим доступу : http://ogo.ua/ articles/view/011–02–23/26490.htm.

16. Остапов С. Е. Технології захисту інформації: навчальний посібник / С.Е. Остапов, С.П. Євсеєв, О.Г. Король – Харків : Вид-во ХНЕУ, 2016. – 476 с.

17. Лєнков, С.В. Аналіз існуючих методів та алгоритмів виявлення атак в бездротових мережах передачі даних / С.В. Лєнков, В.М. Джулій, Н.М. Берназ, С.О. Божук // Збірник наукових праць Військового інституту Київського національного університету імені Тараса Шевченка. – К.: ВІКНУ, 2017. – Вип. № 56. – С.124-132

18. Бурячок В. Л. Інформаційний та кіберпростори: проблеми безпеки, методи та засоби боротьби : посібник / [В. Л. Бурячок, С. В. Толюпа, В. В. Семко та ін.]. – К. : ДУТ-КНУ, 2016. – 178 с.

19. Рибальченко Л.В., Косиченко О.О. Проблеми безпеки персональних даних в Україні / Регіональна економіка / Запоріжжя. 2019. – с.57-62

20. Джулій, В.М. Метод класифікації додатків трафіка комп'ютерних мереж на основі машинного навчання в умовах невизначеності / В.М. Джулій, О.В. Мірошніченко, Л.В. Солодєєва // Збірник наукових праць Військового інституту Київського національного університету імені Тараса Шевченка. – К.: ВІКНУ, 2022. – Вип. №74. – С. 73-82.

21. Лавров, Є. А. Математичні методи дослідження операцій : підручник / Є. А. Лавров, Л. П. Перхун, В. В. Шендрик – Суми : Сумський державний університет, 2017. – 212 с.

22. Гончар С. Ф. Оцінювання ризиків кібербезпеки інформаційних систем об'єктів критичної інфраструктури : монографія. / С. Ф. Гончар. – Київ, 2019. – 175 с.

23. Флах, П. Машинне навчання. Наука та мистецтво побудови алгоритмів, які вилучають знання з даних / П. Флах. — Litres, 2019р.-534с.

24. Хорошко, В.О. Захист систем електронних комунікацій: навч. посіб. / В.О. Хорошко, О.В. Криворучко, М.М. Браіловський - Київ., 2019р. 164 с.

25. Yemchuk L. Organizational Network Analysis as a Tool for Leadership Assessment in Software Development Team. Zhylinska O.; Chornyi A.; Dzhuliy V. – Institute of Electrical and Electronics Engineers (30 September 2020); INSPEC Accession Number: 20008165; DOI: 10.1109/ACIT49673.2020.

26. Сигнатура атаки. Wikipedia [Електронний ресурс] – Режим доступу до ресурсу: https://uk.wikipedia.org/wiki/Сигнатура_атаки.

27. OPWNAI: Cybercriminals Starting to Use ChatGPT, January 6, 2023 [Електронний ресурс] – Режим доступу до ресурсу: https://research.checkpoint.com/2023/opwnai-cybercriminals-starting-to-usechatgpt.

REFERENCES:

1. Doktryna informatsiinoi bezpeky Ukrainy, zatverdzhenoi Ukazom Prezydenta Ukrainy vid 25 liutoho 2017 roku № №47/2017, 15s.

2. Derzhavnyi standart Ukrainy Zakhyst informatsii. Tekhnichnyi zakhyst informatsii. Osnovni polozhennia. DSTU 3396.0-96 [Elektronnyi resurs]. – Rezhym dostupu: http://www.dsszzi.gov.ua/dsszzi/control/uk/publish/article?art_id=38883&cat_id =38836

3. Bem, M. V. (2018) Standarty zakhystu personalnykh danykh v sotsialnii sferi. / M. V.Bem, I. M. Horodyskyi -Lviv - 110 s.

4. Bohush, V.M. (2015) .Informatsiina bezpeka derzhavy / V.M. Bohush, O.K. Yudin. – K.: MK-Pres, – 432 s.

5. Holubiev, O.V. (2018) Prohramno-tekhnichni zasoby zakhystu danykh vid kompiuternykh zlochyniv / O. V. Hoshubiev– Zaporizhzhia : «Pavel» – 145 s.

6. Horbulin, P.V. (2019) Problemy zakhystu informatsiinoho prostoru Ukrainy / M.M. Bachenok, P.V. Horbulin – K.: Intertekhnolohiia – 138 s.

7. Lenkov, S.V.(2023), Metod prohnozuvannia vrazlyvostei informatsiinoi bezpeky na osnovi analizu danykh tematychnykh internet-resursiv / S.V. Lienkov, V.M. Dzhulii, A.M. Bernaz, I.V. Muliar, I.V. Pampukha // Zbirnyk naukovykh prats Viiskovoho instytutu Kyivskoho natsionalnoho universytetu imeni Tarasa Shevchenka. – K.: VIKNU -. №78. – C. 123-134.

8. Lenkov, S.V.(2022) Metod protydii poshyrenniu ta vyiavlennia shkidlyvoi informatsii v sotsialnykh merezhakh/ S.V. Lenkov, V.M. Dzhulii, L.V. Solodieieva // Zbirnyk naukovykh prats Viiskovoho instytutu Kyivskoho natsionalnoho universytetu imeni Tarasa Shevchenka. – K.: VIKNU. – Vyp. №77. – C. 103-117.

9. Lenkov, S.V. (2020), Model bezpeky poshyrennia zaboronenoi informatsii v informatsiino-telekomunikatsiinykh merezhakh / S.V. Lenkov, V.M. Dzhulii, V.S. ORLENKO, O.V. Sieliukov, A.V. Atamaniuk // Zbirnyk naukovykh prats Viiskovoho instytutu Kyivskoho natsionalnoho universytetu imeni Tarasa Shevchenka. – K.: VIKNU. – №68. – pp. 53-64.

10. Dzhulii, V.M. (2023) Alhorytmy prohnozuvannia vrazlyvostei ta zahroz informatsiinoi bezpeky na osnovi tematychnykh internet-resursiv/ Maior Ye., Dzhulii V., Cheshun V., Petliak N. Mizhnarodnyi naukovo-tekhnichnyi zhurnal «Vymiriuvalna ta obchysliuvalna tekhnika v tekhnolohichnykh protsesakh». Vypusk 4. S.49-56.

11. Lienkov, S.V. (2023) Informatsiino-analitychna systemy prohnozuvannia vrazlyvostei ta zahroz informatsiinoi bezpeky/ S.V. Lienkov, V.M. Dzhulii, O.V. Miroshnichenko, V.O. Braun, S.I. Prokhorskyi // Zbirnyk naukovykh prats Viiskovoho instytutu Kyivskoho natsionalnoho universytetu imeni Tarasa Shevchenka. – Kyiv: VIKNU, 2023. – Vyp. № 79. – C. 114-127

12. Yemelianov, S.L. (2019) Osnovy informatsiinoi bezpeky./S.L.Yemelianov– Odesa: Feniks – 357s.

13. Kudinov, V.A. (2016) Osnovy protydii kiberzlochynnosti. / V. M. Smahliuk, V. H. Khakhanovskyi, V.A. Kudinov. – K. : NAVS – 104 s.

14. Lukianov, B. V. (2017) Kompiuternyi analiz danykh / B. V. Lukianov – K. : Akademiia – 345 s.

15. Cotsialni merezhi – realni zahrozy virtualnoho svitu. [Elektronnyi resurs]. – Rezhym dostupu : http://ogo.ua/ articles/view/011– 02–23/26490.htm

16. Ostapov, S. E. (2016) Tekhnolohii zakhystu informatsii: navchalnyi posibnyk / S.E. Ostapov, S.P. Yevseiev, O.H. Korol–Kharkiv : Vyd-vo KhNEU. – 476 s.

17. Lenkov, S.V. (2017), AnalIz Isnuyuchih metodiv ta algoritmiv viyavlennya atak v bezdrotovih merezhah peredachI danih / S.V. Lenkov, V.M. Dzhuliy, N.M. Bernaz, S.O. Bozhuk // Zbirnik naukovih prats Viyskovogo Institutu Kiyivskogo natsionalnogo universitetu imeni Tarasa Shevchenka. – K.: VIKNU. – Vip. No 56. – p.124-132

18. Buriachok, V. L. (2016) Informatsiinyi ta kiberprostory: problemy bezpeky, metody ta zasoby borotby : posibnyk / V. L. Buriachok, S. V. Toliupa, V. V. Semko – K. : DUT-KNU – 178 s.

19. Rybalchenko, L.V., Kosychenko, O.O. (2019) Problemy bezpeky personalnykh danykh v Ukraini / Rehionalna ekonomika / Zaporizhzhia – s.57-62

20. Dzhulii, V.M. (2022), Metod klasyfikatsii dodatkiv trafika kompiuternykh merezh na osnovi mashynnoho navchannia v umovakh nevyznachenosti / V.M. Dzhulii, O.V. Miroshnichenko, L.V. Solodieieva // Zbirnyk naukovykh prats Viiskovoho instytutu Kyivskoho natsionalnoho universytetu imeni Tarasa Shevchenka. – K.: VIKNU. – Vyp. №74. – pp. 73-82.

21. Lavrov, Ye. A. (2017.), Matematychni metody doslidzhennia operatsii : pidruchnyk / Ye. A. Lavrov, L. P. Perkhun, V. V. Shendryk – Sumy : Sumskyi derzhavnyi universytet – 212 p

22. Honchar, S. F. (2019) Otsiniuvannia ryzykiv kiberbezpeky informatsiinykh system obiektiv krytychnoi infrastruktury : monohrafiia. / S. F. Honchar. – Kyiv – 175 s.

23. Flakh, P. Mashynne navchannia. Nauka ta mystetstvo pobudovy alhorytmiv, yaki vyluchaiut znannia z danykh / P. Flakh. — Litres, 2019r.-534s.

24. Khoroshko, V.O. Zakhyst system elektronnykh komunikatsii: navch. posib. / V.O. Khoroshko, O.V. Kryvoruchko, M.M. Brailovskyi - Kyiv., 2019r. 164 s.

25. Yemchuk L. Organizational Network Analysis as a Tool for Leadership Assessment in Software Development Team. Zhylinska O.; Chornyi A.; Dzhuliy V. – Institute of Electrical and Electronics Engineers (30 September 2020); INSPEC Accession Number: 20008165; DOI: 10.1109/ACIT49673.2020.

26. Syhnatura ataky. Wikipedia [Elektronnyi resurs] – Rezhym dostupu do resursu: https://uk.wikipedia.org/wiki/Syhnatura_ataky.

27. OPWNAI: Cybercriminals Starting to Use ChatGPT, January 6, 2023 [Elektronnyi resurs] – Rezhym dostupu do resursu: https://research.checkpoint.com/2023/opwnai-cybercriminals-starting-to-usechatgpt.

Dr. Tech. Sci. Prof. Lienkov S.V., Ph.D Dzhuliy V.M.,
PhD. Mulyar I.V., Ph.D. Kubiavka M.B

FUNCTIONAL MODEL OF CLASSIFICATION OF PSEUDO-RANDOM SEQUENCES OF ENCRYPTED AND COMPRESSED DATA PREVENTION OF LEAKAGE OF CONFIDENTIAL INFORMATION

The task of building a formalized insider model, which can be used both in commercial and public companies, is considered. It is shown that data security threats are characterized by a set of qualitative and quantitative vector indicators, and their formalization requires the application of fuzzy set theory and discrete mathematics. It is shown that it is impossible to use expert traditional assessment methods to determine most of the considered indicators.

To minimize the risk of leakage of confidential information, it is suggested to form groups of employees and calculate the risk of leakage of confidential data for each of them.

The development of a model of pseudo-random sequences will allow us to assess the degree of influence of statistical features extracted from pseudo-random sequences and used in the process of forming a classifier on the accuracy of the classification procedure. The obtained quantitative values of the features will allow to optimize the number of parameters, subject to the required accuracy, to estimate the complexity of the feature removal procedure. On the basis of the simulation results obtained, the identified features of the classifier, it is necessary to justify the choice of a mathematical apparatus, which will allow us to proceed to the practical implementation of the sequence classification algorithm formed by data compression and encryption algorithms.

The conducted analysis of research in this subject area made it possible to identify a practical problem of existing protection mechanisms: low accuracy of detecting encrypted information, due to their similarity to typical high-entropy sequences, use of service information inherent in the transmission process, storage of confidential information. Thus, the task of classifying encrypted and compressed data is relevant.

In order to solve the given task, it is necessary to: conduct an analysis of the features of the functioning of prospective means of preventing and detecting the leakage of confidential data, identify the limitations associated with the detection of compressed and encrypted information, justify the choice of an appropriate feature space for modeling pseudo-random sequences formed by information compression and encryption algorithms; to develop a model of pseudo-random sequences formed by data compression and encryption algorithms, which differs from known ones, taking into account their statistical characteristics.

The presented model of pseudorandom sequences differs from analogs taking into account the distribution of bytes and taking into account the frequencies of bit subsequences of length 9 bits. To assess the adequacy of the proposed model, experiments were conducted to determine the accuracy of classification of pseudorandom sequences by machine learning algorithms.

Keywords: pseudorandom sequences, functional model, information security, classification accuracy, encrypted, compressed data.