Хмельницький національний університет
Факультет інформаційних технологій
Кафедра кібербезпеки

КВАЛІФІКАЦІЙНА РОБОТА МАГІСТРА

на тему
Метод захисту від витоку інформації на основі поділу стислих та зашифрованих даних

Галузь знань 12 – Інформаційні технології

Спеціальність 125 –Кібербезпека

КРМКБ. 2019/035.22.01.14 ПЗ

Виконав: студент 2 курсу, група КБм-22-1 _____ Кучерявий Є.І.
Підпис

Керівник доц., к. т. н, доцент _____ Джулій В.М.
Підпис

Нормоконтролер старший викладач _____ Мостовий С.В.
Підпис

До захисту допускаю:
Зав. кафедри кібербезпеки, к.т.н., доц _____ Кльоц Ю.П.
Підпис

11 12 2023р.

Хмельницький, 2023

ХМЕЛЬНИЦЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ

Факультет ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ

Кафедра КІБЕРБЕЗПЕКИ

Освітній рівень МАГІСТР

Галузь знань 12 ІНФОРМАЦІЙНИХ ТЕХНОЛОГІЙ

Спеціальність 125 КІБЕРБЕЗПЕКА

Освітня програма КІБЕРБЕЗПЕКА

ЗАТВЕРДЖУЮ

Зав. кафедри Ю.П. Кльоц

"30" 08 2023 р.

ЗАВДАННЯ
НА КВАЛІФІКАЦІЙНУ РОБОТУ

Кучерявому Євгену Ігоровичу
(прізвище, ім'я, по батькові)

1. Тема проекту (роботи) Метод захисту від витоку інформації на основі поділу стислих та зашифрованих даних

Науковий керівник Джулій Володимир Миколайович, к.т.н., доцент
(прізвище, ім'я, по батькові, науковий ступінь, вчене звання)

Затверджена наказом № 30 ректора університету, додаток №25 від 15.08.2023

2. Строк подання студентом проекту (роботи) на кафедру 05.12.2023.

3. Вихідні дані до проекту (роботи) Провести аналіз особливостей функціонування засобів запобігання та виявлення витоку конфіденційних даних. Розробити модель, сформованих алгоритмами стиснення та шифрування даних, ПВП. Розробити метод класифікації ПВП, який враховує статистичні ознаки послідовностей.

4. Зміст пояснювальної записки (перелік питань, які потрібно розробити)
Дослідження сучасного стану стану предметної області витоку інформації на основі поділу стислих та зашифрованих даних. Функціональна модель класифікації псевдовипадкових послідовностей. Модель псевдовипадкових послідовностей з врахуванням їх статистичних характеристик. Метод класифікації псевдовипадкових послідовностей сформованих алгоритмами стиснення та шифрування інформації.

5. Перелік графічного матеріалу (із зазначенням обов'язкових креслень)

6. Консультанти розділів дипломного проекту (роботи)

Розділ	Прізвище, ініціали і посада консультанта	Підпис, дата	
		завдання видав	завдання прийняв
Нормоконтроль	Мостовий С.В. Старший викладач кафедри кібербезпеки		

7. Дата видачі завдання: «01»___лютого_____2023_р.

КАЛЕНДАРНИЙ ПЛАН

№ з/п	Назва етапів дипломної роботи	Строк виконання етапів роботи	*Примітка*
1	Ґрунтовне ознайомлення та дослідження предметної галузі	21.02.2023	Виконано
2	Визначення змісту, структури магістерської роботи	11.03. 2023	Виконано
3	Опрацювання першого розділу магістерської роботи	4.04. 2023	Виконано
4	Опрацювання статті за результатами дослідження	3.05. 2023	Виконано
5	Опрацювання другого розділу магістерської роботи	2.06. 2023	Виконано
6	Опрацювання третього розділу магістерської роботи	2.09. 2023	Виконано
7	Опрацювання четвертого розділу магістерської роботи	4.10. 2023	Виконано
8	Підготовка та опрацювання ілюстративного матеріалу	7.11. 2023	Виконано
9	Оформлення магістерської роботи графічної та текстової частини	18.11. 2023	Виконано
10	Попередній захист магістерської роботи	25.11. 2023	Виконано
11	Захист магістерської роботи на засіданні ЕК	5.12. 2023	Виконано

Студент _____ Є.І. Кучерявий
Підпис Ініціали, прізвище

Керівник проекту (роботи) _____ В.М. Джулій
Підпис Ініціали, прізвище

АНОТАЦІЯ

Тема кваліфікаційної роботи: «Метод захисту від витоку інформації на основі поділу стислих та зашифрованих даних».

Автор роботи: Кучерявий Євген Ігорович

Керівник роботи: к.т.н. доц. Джулій Володимир Миколайович

Загальний обсяг роботи: 82 сторінки, 26 рисунків, 9 таблиць, 4 додатки, 58 посилань.

Ключові слова: моделі, алгоритми, псевдовипадкові послідовності, методи класифікації, шифрування інформації.

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.

В роботі представлено:

1. Модель псевдовипадкових послідовностей, сформованих алгоритмами стиснення та шифрування даних, дозволила врахувати особливості стиснених та зашифрованих псевдовипадкових послідовностей при поданні в бінарному виді, підпослідовностями довжиною в дев'ять біт.

2. Метод класифікації псевдовипадкових послідовностей, сформованих алгоритмами стиснення та шифрування даних, враховує дискримінуючу здатність статистичних ознак послідовностей, показує більш високу точність класифікації на відміну від відомих аналогів.

Проведено оцінку ефективності запропонованих підходів. Отримані значення точності класифікації псевдовипадкових послідовностей перевищують відомі аналоги.

11.12.2023

ANNOTATION

Theme of qualification work: "A method of protection against information leakage based on the division of compressed and encrypted data".

Author of the work: Kucheryavy Evgeny Ihorovych

Mentor: Ph.D. Dguliy Volodymyr Mykolayovych

Total volume of work: 82 pages, 26 figures, 9 tables, 4 appendices, 58 links.

Keywords: models, algorithms, pseudorandom sequences, classification methods, information encryption.

The analysis of information security incidents carried out by the analytical centers of SafeNet companies shows that in cases of leakage of confidential data, more than 52% of the culprits were internal violators.

The work presents:

1. The model of pseudo-random sequences formed by data compression and encryption algorithms made it possible to take into account the features of compressed and encrypted pseudo-random sequences when presented in binary form, sub-sequences nine bits long.

2. The method of classification of pseudo-random sequences formed by data compression and encryption algorithms takes into account the discriminating ability of statistical features of sequences, shows a higher accuracy of classification in contrast to known analogues.

The effectiveness of the proposed approaches was evaluated. The obtained accuracy values of the classification of pseudorandom sequences exceed the known analogues.

11.12.2023

ЗМІСТ

ВСТУП

Інформаційні технології, на теперішній час, 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 [1,4,33].

Незважаючи на удосконалення механізмів захисту від кіберзагроз, розвиток засобів захисту конфіденційної даних, 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, але здійснювати передачу конфіденційної інформації за контрольований інформаційний периметр компанії [1,5,9,11,19].

Забезпечення інформаційної безпеки конфіденційних даних та призупинення дій внутрішнього порушника здійснюється за допомогою, в основному, організаційних заходів. Виконується тестування, перевірка фактів їхньої біографії, відбір кандидатів, протягом проміжку часу можуть змінитися багато факторів, один з яких - лояльність співробітника [2,4,9,13,14].

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 [30-34].

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 [37,39,42,51,58].

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.

Об'єктом дослідження: псевдовипадкові послідовності, сформовані алгоритмами стиснення та шифрування інформації.

Предмет дослідження: алгоритми, методи класифікації, сформованих алгоритмами стиснення та шифрування інформації, псевдовипадкових послідовностей.

Мета магістерської роботи: підвищення точності класифікації, сформованих алгоритмами стиснення та шифрування інформації, псевдовипадкових послідовностей.

Наукове задача: розробка методу класифікації, сформованих алгоритмами стиснення та шифрування інформації, псевдовипадкових послідовностей, для захисту від витоку конфіденційних даних в зашифрованому вигляді.

Для досягнення поставленої мети в магістерській роботі необхідно вирішити наступні задачі: 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; розробити метод класифікації, сформованих алгоритмами стиснення та шифрування даних, псевдовипадкових послідовностей, що враховує здатність їх статистичних ознак.

Наукова новизна дослідження:

1. Модель, сформованих алгоритмами стиснення та шифрування інформації, псевдовипадкових послідовностей, відрізняється врахуванням статистичних характеристик послідовностей;

2. Метод класифікації псевдовипадкових послідовностей, сформованих алгоритмами шифрування та стиснення інформації, враховує статистичні ознаки послідовностей.

3. Спосіб класифікації, сформованих алгоритмами стиснення та шифрування інформації, псевдовипадкових послідовностей, для захисту від витоку конфіденційних даних в зашифрованому вигляді.

Практична цінність магістерського дослідження полягає у підвищенні точності класифікації, сформованих алгоритмами стиснення та шифрування інформації, псевдовипадкових послідовностей, для захисту від витоку конфіденційних даних в зашифрованому представленні та відмові від контекстних ознак.

Методи дослідження. У ході проведення магістерського дослідження використано методи теорії розпізнавання образів, математичної статистики, математичного моделювання.

Основні результати магістерського дослідження, що виносяться на захист:

1. Модель, сформованих алгоритмами стиснення та шифрування інформації, псевдовипадкових послідовностей, відрізняється врахуванням статичних характеристик послідовностей.

2. Метод класифікації, сформованих алгоритмами стиснення та шифрування інформації, псевдовипадкових послідовностей, враховує статистичні ознаки послідовностей.

3. Спосіб класифікації, сформованих алгоритмами стиснення та шифрування інформації, псевдовипадкових послідовностей, для захисту від витоку конфіденційних даних в зашифрованому вигляді.

Достовірність отриманих результатів у магістерській роботі підтверджується застосуванням апробованих математичних моделей, коректним

використанням математичного апарату, позитивними результатами основних положень роботи на науково-технічних конференціях, результатами експериментальних досліджень.

Особистий внесок. Дослідження, викладені в магістерській роботі, проведені автором при виконанні дослідження предметної області в процесі наукової діяльності. Результати роботи, які виносяться на захист, отримані автором особисто, запозичений матеріал, використаний в роботі, позначений посиланнями.

Апробація роботи. За темою дипломної роботи ОКР «Магістр» опубліковано 1 теза доповідей, 1 фахова стаття.

Структура і обсяг роботи. Дипломна робота ОКР «Магістр» складається зі вступу, основної частини, що містить 4 розділи, висновків і списку використаних джерел. Загальний обсяг роботи - 82 сторінки. Робота містить 26 рисунків та 9 таблиць. Список використаної літератури включає 58 бібліографічних джерел.

1 АНАЛІЗ СТАНУ ПРЕДМЕТНОЇ ОБЛАСТІ ВИТОКУ ІНФОРМАЦІЇ НА ОСНОВІ ПОДІЛУ СТИСЛИХ ТА ЗАШИФРОВАНИХ ДАНИХ

1.1 Дослідження реалізації загроз витоку конфіденційних даних в корпоративних мережах

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 (внутрішньому порушнику) слабо розвинені, в більшості документів, що регламентують політику безпеки кофіденційним даним компанії, містяться постулати про відсутність інсайдера, що тягне, в даному випадку, зростання ймовірності порушення інформаційної безпеки даних, що захищаються [20,39,41].

Відповідно до звіту міжнародного експертно-аналітичного центру компаній Group-IB частка інсайдер (внутрішніх порушників), як джерел зареєстрованих випадків в організаціях витоку конфіденційної інформації, за період із січня по червень 2022 р. склала понад 80%. У 78% зареєстрованих випадках витоку даних було організовано навмисне [6,11,26,30,31].

Типовими внутрішніми порушниками є співробітники (рядові), 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 [13-15,18,19,22].

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 [4] 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 [29,31,38,43].

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 [27,37,38,58].

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 [7-9,31]. 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 [34-37,46,47].

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 [9]. 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 [5,6,33,34,37]. Відзначається можливість передачі внутрішніми

порушниками захищених даних з контрольованого периметра компанії з використанням сервісів електронної пошти. 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 конфіденційних даних можуть бути політичні, індивідуальні, фінансові мотиви працівників компанії [22,48,50].

Аналіз досліджень мережевої активності корпоративної мережі є ключовим компонентом запобігання та раннього виявлення загроз та атак безпеки конфіденційним даним, що виходять від інсайдерів [1,6,7,12]. 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. Пропонується використовувати нейронні мережі, що дозволить у реальному часі виявляти загрози та атаки. Розроблена модель визначає сумарне значення ймовірності настання атаки, загрози конфіденційним даним з використанням оцінки аномальної поведінки користувача процесу та системи. Для оцінки точності використовується метрика Recall, найбільш повно оцінює виявлення будь-якого класу даних.

Безпека даних (інформації) - стан захищеності даних, при якому забезпечені їх цілісність, доступність, конфіденційність. Витікання інформації - неконтрольоване поширення даних, що захищаються, в результаті несанкціонованого доступу, розголошення та отримання даних, що захищаються іноземними розвідками, несанкціонований доступ до даних третіми особами, організований ненавмисно або навмисно робітниками компанії [7,10,13,17].

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 [22,26].

Для запобігання реалізації атак та загроз витоку конфіденційної інформації в корпоративних мережах застосовують засоби запобігання та виявлення витоку даних (DLP–системи) [30,31,50,57], 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.

Відсутність у корпоративній моделях 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 [1-3,7].

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 [15]. Атаки на такі критичні інформаційні системи включають проникнення в мережу організації та встановлення шкідливого програмного забезпечення, які можуть розкрити конфіденційну інформацію, змінити поведінку конкретного технічного обладнання. 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. 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 [22,23,26-28].

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

1.2 Аналіз методів протидії загрозам витоку інформації з корпоративних мереж

Корпоративні системи зберігання та обробки даних призначені для виконання процедур зберігання, передачі, перетворення інформації різного ступеня секретності, типу, а також персональні дані співробітників, користувачів компанії, а у випадку державної установи також дані великої кількості громадян. Даний факт дозволяє розглядати корпоративні системи компанії, як інформаційні

системи зберігання, обробки персональних даних. У зазначених інформаційних системах під зловмисником розуміється фізична особа, яка навмисно чи випадково вчиняє дії, які є наслідком порушення безпеки персональних даних співробітників при обробці інформації технічними засобами в інформаційно-пошукових системах персональних даних [4,6,17,30,32,36].

Залежно від наявності у зловмисників легітимного доступу до інформаційно-пошукової системи, системи поділяються: зловмисники, які мають доступ до корпоративної мережі організації передачі даних, включаючи, також, користувачів інформаційно-пошукової системи, що реалізують загрози безпосередньо у системі – внутрішні зловмисники; зловмисники, які не мають доступу до корпоративної мережі компанії передачі даних, реалізують атаки, загрози із мереж міжнародного інформаційного обміну або зовнішніх мереж зв'язку загального користування – зовнішні порушники.

Можливості внутрішнього зловмисника на теперішній час дуже великі і становлять серйозну безпеку для конфіденційної інформації, враховуючи, при цьому, широкий спектр програмнно-апаратний засобів, що реалізують процедури стиснення чи шифрування інформації. Внутрішні потенційні зловмисники можуть бути розділені на вісім категорій залежно від повноважень та способу доступу до конфіденційних даних:

– перша категорія - співробітники компанії, що мають легітимний доступ до даних, але не мають доступу до конфіденційної інформації. До цього типу зловмисників належать посадові особи, які забезпечують функціонування інформаційно-пошукових систем;

–друга категорія - користувачі інформаційно-пошукової системи, які мають обмежений доступ із робочого місця до ресурсів організації.

–третя категорія - користувачі інформаційно-пошукової системи, які здійснюють віддалений доступ до ресурсів організації.

–четверта категорія - користувачі інформаційно-пошукової системи з повноваженнями адміністратора безпекиданих підсистеми інформаційної системи.

–п'ята категорія - користувачі інформаційно-пошукової системи з відповідними повноваженнями системного адміністратора.

–шоста категорія - користувачі інформаційно-пошукової системи з повноваженнями адміністратора безпеки даних.

–сьома категорії - постачальники (програмісти-розробники) прикладного ПЗ та особи, які забезпечують супровід ППЗ.

–восьма категорія - особи та розробники, які забезпечують супровід, постачання, ремонт технічних засобів, що реалізують функціонування інформаційно-пошукової системи.

DLP–системи дозволяють проводити аналіз переданої інформації для всіх груп у використання клієнт-серверної архітектури інформаційно-пошукової системи в корпоративній мережі організації, у випадку якщо порушник не використовує методи стиснення та шифрування переданих даних. В даному випадку внутрішній зловмисник здатний реалізувати витік конфіденційної інформації в обхід існуючих політик безпеки та систем захисту, що унеможливлює реєстрацію події безпеки інформації, і також, розслідування інциденту безпеки даних.

Вразливість інформаційно-пошукової системи в корпоративній мережі організації передачі даних є слабким місцем (недоліком) у прикладному, системному програмному забезпеченні автоматизованої інформаційно-пошукової системи, яка використовується для реалізації атак, загроз безпеці конфіденційних даних. Причинами виникнення вразливості є [6,7,20,42,45]: навмисні дії внесення вразливості в ході розробки та проектування програмно-апаратного забезпечення; помилки при розробці та проектуванні програмно апаратного забезпечення; несанкціоноване використання та впровадження неврахованих програм з подальшим витрачанням ресурсів (захоплення оперативної пам'яті, завантаження процесора); неправильні налаштування програмно-апаратного забезпечення, неправомірна зміна режимів роботи програм та пристроїв; несанкціоновані ненавмисні дії співробітників, що призводять до виникнення в системі

вразливостей; збої в роботі програмно-апаратного забезпечення (викликані виходом з ладу апаратних елементів, збоями в електроживленні, зовнішніми впливами електромагнітних полів); впровадження шкідливого програмного забезпечення, що створюють вразливості у програмно-апаратному забезпеченні.

Під вразливістю інформаційно-пошукової системи в рамках корпоративної мережі організації передачі даних розуміється можливість передачі стислої чи зашифрованої інформації, у разі відсутності даних про алгоритм шифрування (стиснення). Вразливість реалізується шляхом можливості внутрішніх зловмисників легітимно використовувати програмно-апаратне забезпечення, що реалізує видозміну даних за рахунок використання процесів стиснення чи шифрування інформації та подолання механізмів захисту DLP–систем.

Для зазначених атак, загроз інформаційній безпеці необхідно визначити значення відповідних показників рівня небезпеки. Відповідно до [18,47,53], показники рівня небезпек даних представлені у табл. 1.1. Внутрішній зловмисник має локальний доступ до інформації організації, оскільки має доступ до мережі, також, можливе впровадження шкідливого програмного забезпечення, яке приховано функціонує в корпоративній мережі організації (botnet агенти).

Таблиця 1.1 - Значення рівня показників небезпеки інформації

Показник рівня небезпеки інформації	Найменування показника небезпеки	Значення показників рівня небезпеки
Тип доступу (td)	Віддалений	3
	Фізичний	1
	Локальний	2
Рівень складності (dl)	Помірний	4
	Середній	2
	Підвищений	3
	Високий	1
Значимість інформаційних компонентів, ресурсів (rc)	Низька	1
	Середня	2
	Висока	3

Рівень складності реалізації небезпеки даних є помірним (найнижчим) з представлених, що обумовлюється простою схемою реалізації загрози передачі даних в стислому чи зашифрованому виді як внутрішнім зловмисником так і шкідливим програмним забезпеченням. Показник рівня небезпеки загрози даним складається із суми значень показників і становить 8, відповідно до виразу (1.1):

$$(W) = td(2) + dl(4) + rc(2) = 8 \quad (1.1)$$

Відповідно до методики [18], визначимо рівень загроз безпеки за табл. 1.2.

Таблиця 1.2 - Рівень небезпеки загроз безпеки даних

Рівень загроз безпеки даних	Діапазон значень
Низький	$W = 4$
Середній	$5 \le W \le 7$
Високий	$W = 8$

Значення отримане у виразі (1.1) відповідає високому рівню загрози інформаційній безпеці даних, що підтверджує, в даній ситуації, актуальність теми досліджень. На рис. 1.1 представлена схема процесу витоку даних в стислому або зашифрованому вигляді за контрольований периметр компанії.

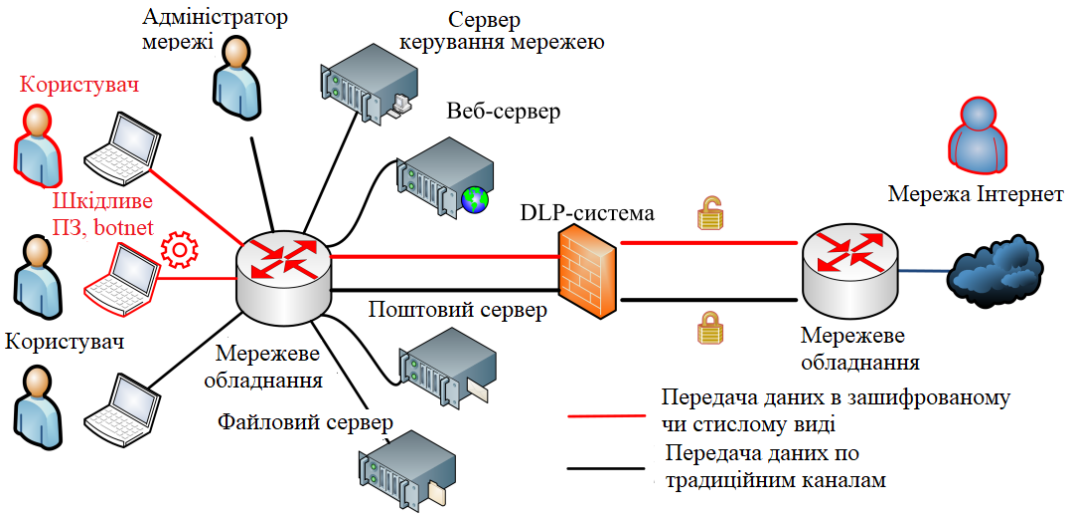


Рисунок 1.1 - Схема процесу витоку даних, реалізована внутрішнім порушником

Внутрішній порушником (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 та шифрування інформації. У разі використання засобів стиснення та 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. Приховані канали передачі інформації, також з використанням методів стеганографії та знімання даних з електромагнітних, побічних оптичних видів каналів у магістерській роботі не розглядаються.

1.3 Аналіз методів класифікації стиснутих та зашифрованих даних засобами запобігання та виявлення витоку інформації

З метою обґрунтування актуальності завдання класифікації зашифрованих, стислих та відкритих даних було проведено аналіз досліджень у галузі захисту інформації. За даними експертно-аналітичного центру SafeNet в 2021 р. в Україні близько 78% зафіксованих витоків конфіденційних даних сталися з вини внутрішніх зловмисників, близько 76% з них були навмисними [1]. Статистика зафіксованих витоків інформації з джерела за 2021 р. представлена на рис. 1.2.

Загрози від внутрішніх зловмисників є важкоусувними та найбільш небезпечними, у тому числі при використанні засобів стиснення та шифрування інформації, що дозволяє, при цьому, не порушуючи політик безпеки підприємства відправляти конфіденційні дані за периметр організації. Ексфільтрація конфіденційної інформації за периметр компанії (крадіжка даних) може здійснюватися широким колом користувачів, що належать до внутрішніх зловмисників. Найбільша кількість витоків конфіденційної інформації

спостерігається в великих організаціях в корпоративних мережах, що здійснюють обробку персональної інформації, а також з мереж державних установ та організацій. Основними джерелами витоків інформації - системи та сервіси електронної пошти, що мають доступ до мережі Інтернет.

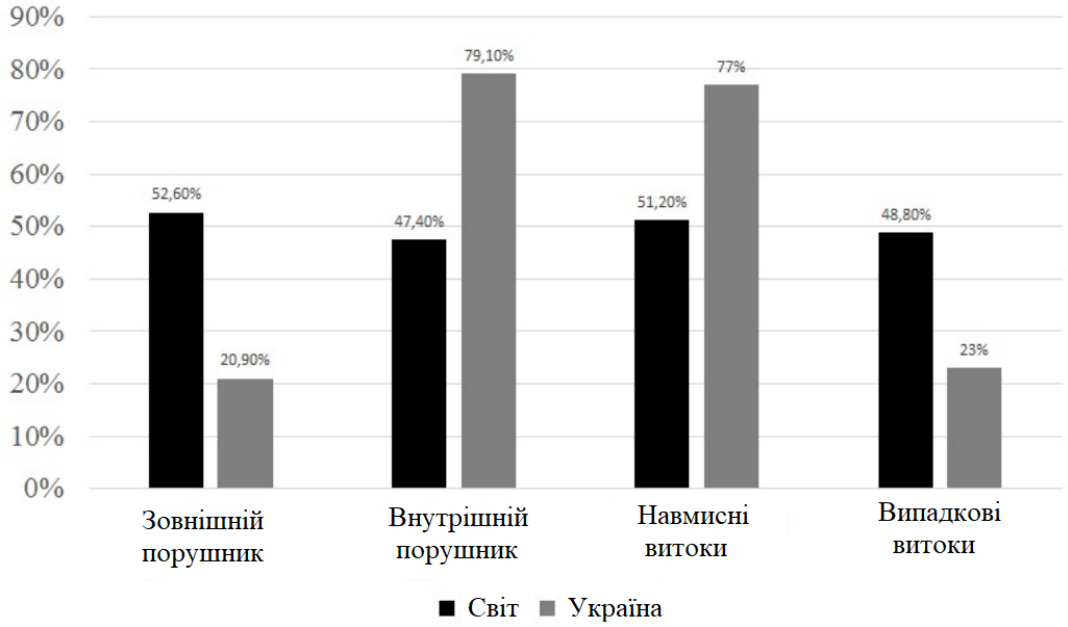


Рисунок 1.2 - Статистика зафіксованих витоків інформації за 2021р.

Високий відсоток витоків конфіденційної інформації може бути обумовлений наявністю недоліків та вразливості у існуючих засобах захисту даних, також в DLP-системах. Сучасні засоби захисту даних не виявляють канали витоку інформації, сформованою шляхом використання шифрування даних, що робить такий захист малоефективними забезпеченням конфіденційності даних.

Методи, що розробляються, класифікації стислих і зашифрованих даних, повинні забезпечити підвищення точність класифікації, застосовуватися в архітектурі клієнт-сервер на серверній стороні. Даний підхід дозволить скоротити час, що витрачається на перевірку послідовностей, скоротити обчислювальні ресурси, необхідні для проведення аналізу даних. Підсистема захисту даних від витоку інформації повинна бути інваріантною щодо форматів представлення даних та будь-якої іншої службової інформації. Незалежно від контейнера

інформації, що передається за периметр підприємства, повинен визначатися тип даних, стислий або зашифрований. Даний підхід дозволить у режимі реального часу проводити класифікацію потенційно небезпечної інформації та своєчасно реагувати на інциденти безпеки інформації.

Для виявлення передачі стиснутих чи зашифрованих даних, які містять конфіденційну інформацію та визначення джерела несанкціонованого поширення даних, необхідно розробити алгоритм класифікації псевдовипадкових послідовностей. Псевдовипадкові послідовності - файли, що проходять тести NIST на випадковість [31], розподіл байт в яких підпорядковується рівномірному закону розподілу, їх ентропія має значення більше 7,5. Для вирішення даної задачі необхідно провести порівняльний аналіз засобів, методів, технологій класифікації відкритих, стислих, зашифрованих даних.

Для запобігання витокам конфіденційних даних застосовують технічні та організаційні заходи, які різняться застосуванням апаратних та програмних засобів. Найбільш поширеним програмним засобом запобігання витоку конфіденційної інформації є DLP-системи, що здійснюють аналіз потоків даних на предмет наявності конфіденційної інформації [22-24,26].

Зазначені методи виконують аналіз потоків інформації на предмет наявності фраз, певних слів, регулярних виразів, здійснюють оцінку контексту переданих даних, способів і службових характеристик протоколів передачі інформації. Однак, існують засоби обходу подібних механізмів захисту, наприклад стиснення або шифрування даних [28].

Класифікація методів, які застосовують DLP-системи для виявлення конфіденційної інформації, представлена на рис. 1.3. Методи можна розділити на дві групи: контекстні та контентні. Контекстні методи - орієнтовані на конкретні протоколи передачі інформації та технології, враховують різні ознаки та факти, що супроводжують процеси обміну даними, адреси одержувачів, джерел, розмір пакета, номер порту програмного забезпечення, що здійснює передачу. Контентні

методи здійснюють пошук зліпків, цифрових відбитків, регулярних виразів, здійснюють аналіз переданої інформації, також службової інформації.

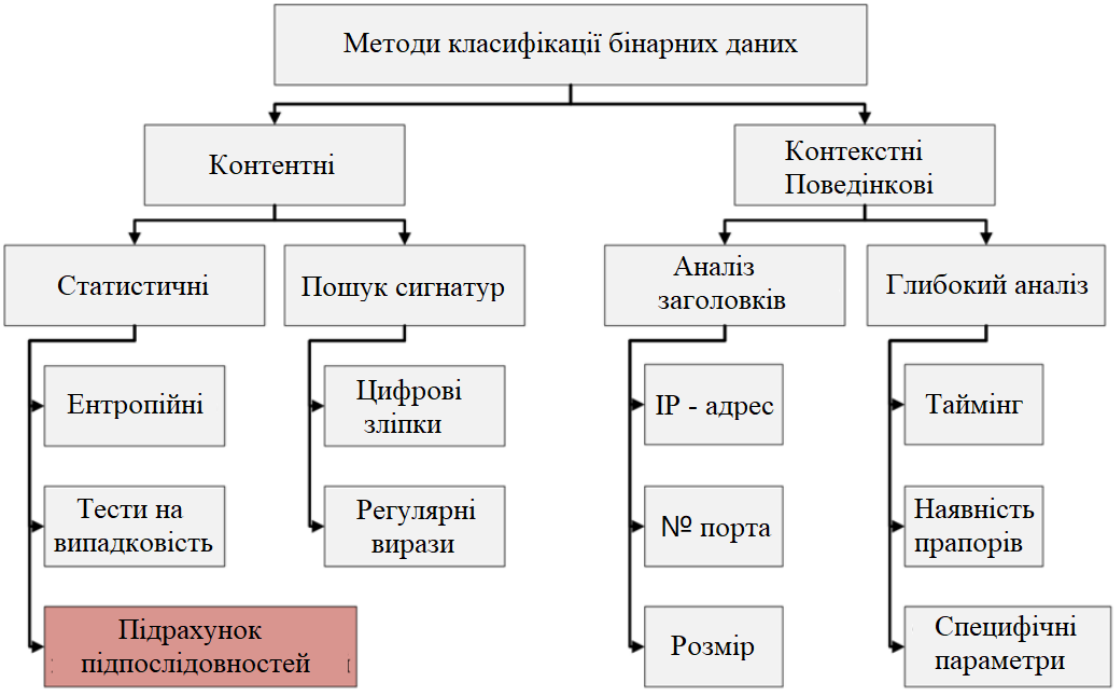


Рисунок 1.3 - Класифікація методів, що використовуються в DLP-системах

До контентних методів відносяться ентропійні підходи - виконують підрахунок ентропії блоків інформаціїх різної довжини, проте не застосовні для класифікації стислих і зашифрованих даних. Статистичні методи і тести на випадковість становлять інтерес для досліджень, так як до алгоритмів шифрування пред'являються певні стандарти по розсіюванню вихідних даних, до алгоритмів стиснення подібні вимоги не висуваються.

Оскільки співробітник підприємства має можливість шифрувати конфіденційну інформацію та передати поза периметр організації, задача ідентифікації стислих та шифрованих даних є актуальною. Зі зростанням популярності криптографічних протоколів та їх впровадженні в телекомунікаційні Інтернет мережі деякі засоби безпеки даних, засновані на глибокому аналізі пакетів, перестають достовірно працювати, не можуть виділити ознаки, на яких здійснювалося детектування потенційно небезпечні дії. Відзначається, також

збільшену складність проведення аналізу бінарних файлів, з огляду на збільшення їх кількості, поділяють бінарні послідовності на три класи: з низькою ентропією (нестислі медіа-файли), середньою ентропією (структури даних, текст, виконувані файли), високоентропійні – зашифровані та стислі дані. Для класифікації з трьох класів послідовностей пропонують використовувати наступні ознаки: ентропія Шеннона, середнє значення байт у послідовності, вага Хеммінга, ψ -квадрат.

Системи, що здійснюють класифікацію інформації, дозволяють зробити перший крок до виявлення шкідливих дій та вторгнень користувачів систем. Спочатку системи захисту класифікувалися двома способами: на основі глибокого аналізу властивостей IP- пакета без аналізу даних і на основі аналізу заголовка пакета (номер порту, IP-адреса). Пропонується використовувати на основі дерева рішень методи машинного навчання, генетичних алгоритмів та адаптивного бустингу. Отримані результати свідчать про можливість класифікації відкритих (незашифрованих) та зашифрованих даних з точністю більше 0,96.

Проведений аналіз досліджень предметної області та об'єкта дослідження дозволяє висунути припущення про наявність у стислих та зашифрованих даних статистичних особливостей. У разі справедливості висунутої гіпотези в результаті проведених досліджень можливо реалізувати модель ПВП, сформованих алгоритмами шифрування інформації і запропонувати метод захисту від витоків переданої інформації на основі поділу типів даних.

1.4 Постановка задачі

Розглянуті методи класифікації переданих стислих і зашифрованих даних дозволяють висунути вимоги, яким повинні задовольняти засоби захисту запобігання та виявлення витоку конфіденційної інформації: безперервність роботи системи в часі; оперативність аналізу переданих даних; незалежність від типу контейнера даних.

До методу безпеки даних від витоку даних, пред'являються наступні вимоги: використання статистичних методів, незалежних від характеристик контейнерів передачі та зберігання даних; класифікація лише підозрілих послідовностей; формат аналізованих даних не важливий, дані надходять в бінарному вигляді; точність класифікації стислих та зашифрованих даних має досягати максимального значення; можливість протидії загрозам безпеці корпоративних мереж підприємства, також протидія та виявлення botnet мережам; час виконання класифікації має досягати до мінімуму.

Таким чином, задача у формальному вигляді визначена виразом (1.2):

$$\begin{aligned} & F\left(x_i, y_i\right)=1, i=j, \\ & i, j \in Y=(0,1,2), \\ & t \rightarrow \min, \\ & Accuracy \rightarrow \max, \end{aligned} \tag{1.2}$$

де x_i – файл, що аналізується, y_i - клас файла x_i, $i, j \in Y$ – класи даних: відкриті, зашифровані, стислі.

Здійснено формальну постановку задачі, визначено мету, проведено аналіз об'єкта та предмета дослідження. Відсоток інцидентів порушення безпеки, конфіденційних даних, пов'язаних із витоком інформації, причиною яких є внутрішні зловмисникм, склав понад 78%, що підтверджує актуальність дослідження. Проведено аналіз вразливостей та загроз DLP–систем та засобів захисту даних, визначено недоліки та переваги використовуваних підходів класифікації стислих та зашифрованих даних.

Обґрунтовано вибір статистичних методів проведення аналізу переданих даних для побудови класифікатора, сформульована наукова задача магістерського дослідження у формальному виді. Показано практичну проблему, яка полягає в низькій точності класифікації стислих і зашифрованих псевдовипадкових послідовностей дорівнює 0,95 та використання заголовків зазначених файлів.

2 ФУНКЦІОНАЛЬНА МОДЕЛЬ КЛАСИФІКАЦІЇ ПСЕВДО-ВИПАДКОВИХ ПОСЛІДОВНОСТЕЙ

2.1 Аналіз результатів дослідження предметної області, ознак, що описують зашифровані і стиснуті послідовності

Для розробки моделі псевдовипадкових чисел необхідно розглянути результати аналізу досліджень предметної області, визначити які ознаки найчастіше використовувані, що описують стислі і зашифровані послідовності. Розробка моделі послідовностей дозволить оцінити ступінь впливу статистичних ознак, що витягуються з псевдовипадкових послідовностей і використовуються, надалі в процесі формування класифікатора, на точність поведення 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 і шифрування даних.

Проведений аналіз відкритого та зашифрованого трафіку на основі підрахунку ентропії окремих слів довжиною 2..64біт, 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 даних 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 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 мережі Tor - 0,989.

У роботі [37] автори 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 [33]. 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 [34], 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 [35] для вирівнювання статистики розподілу слів у досліджуємих файлах. Для побудови класифікатора потоку даних використовувалися алгоритми машинного навчання та отрамані 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%.

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.

2.2 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.

Функціональна модель формування класифікатора представлена на рис. 2.1. 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.

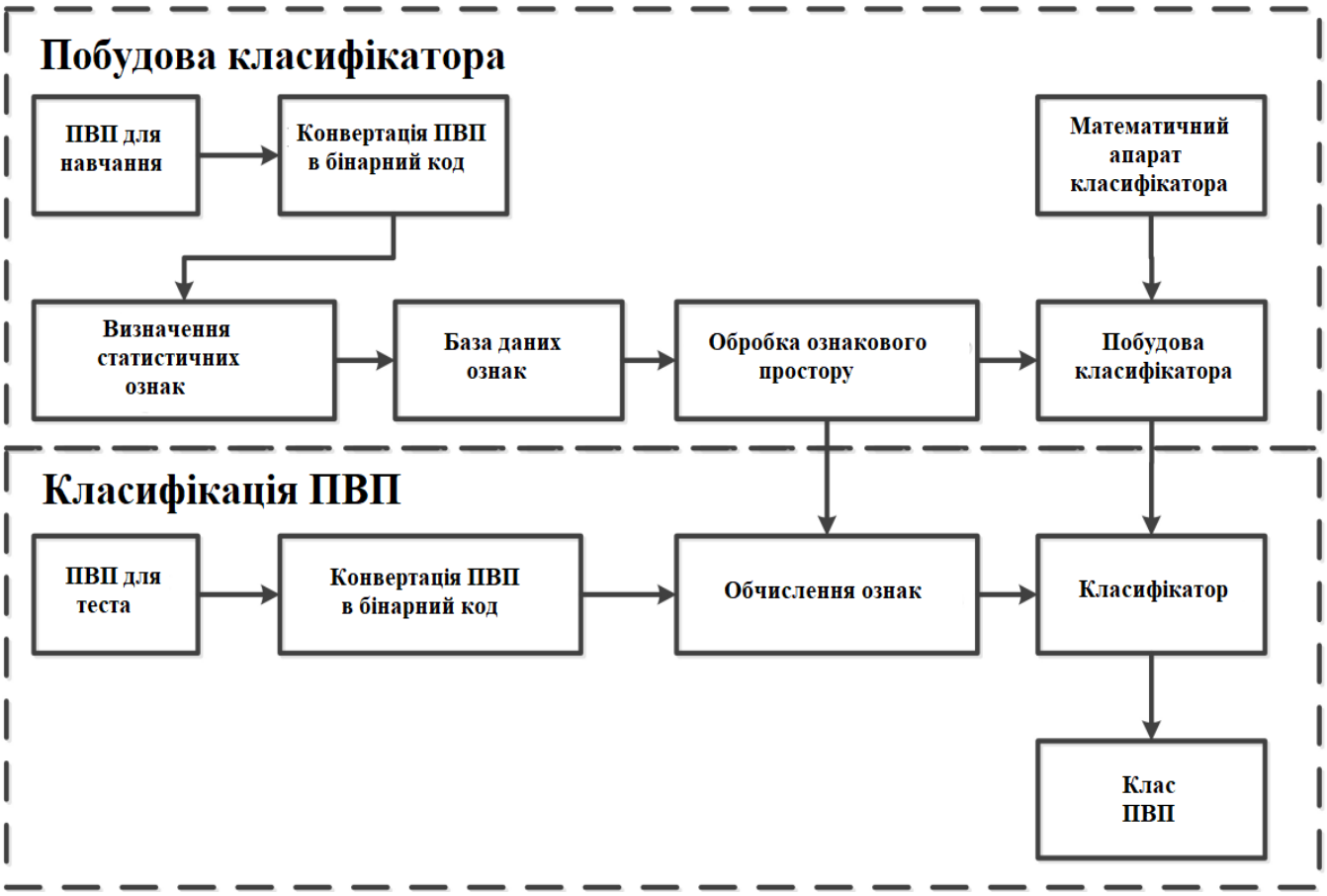


Рисунок 2.1 - Функціональна модель процесу формування класифікатора

Для розподілу вхідної вибірки псевдовипадкової послідовності застосовується процедура стратифікованого вибору груп, що дозволяє отримати різні набори потоку даних для проведення навчання класифікатора (рис. 2.2).

Рисунок 2.2 - 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 [30-32]. 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. 2.3. На рис. 2.3 наведена графічна інтерпретація метрик Recall (повнота) і Precision (точність).

Як наведено на рис. 2.3 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.

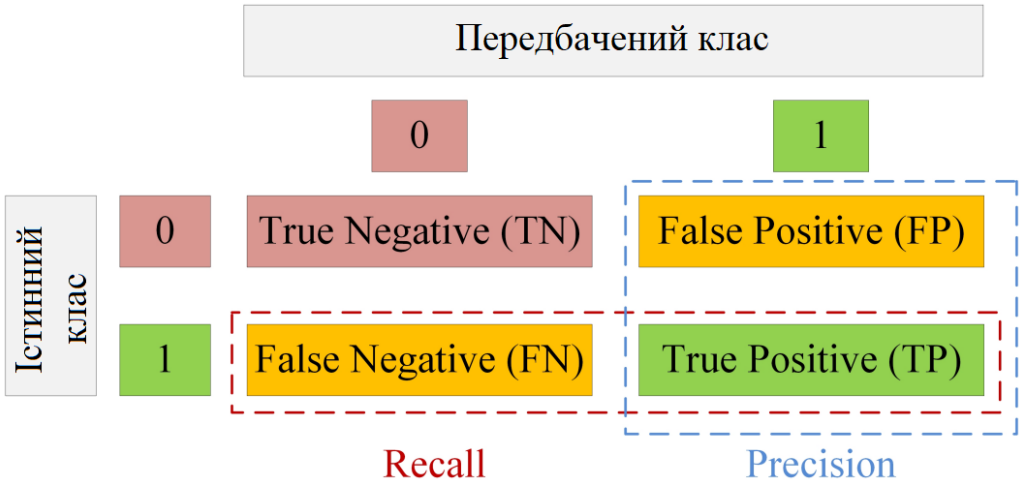


Рисунок 2.3 - Матриця помилок під час проведення бінарної класифікації псевдовипадкових послідовностей

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. 2.4. 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 називають помилкою II-го роду, другий - помилкою I-го роду.

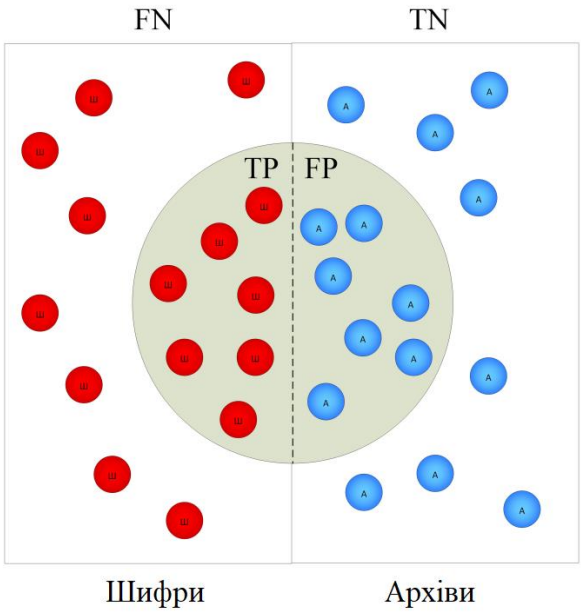


Рисунок 2.4 - Приклад класифікації псевдовипадкових послідовностей

Метрики точність, повнота, частка правильних відповідей, у формальному виді представлені виразом (2.1):

$$\begin{aligned} Precision &= \frac{TP}{TP + FP}, \\ Recall &= \frac{TP}{TP + FN}, \\ Accuracy &= \frac{TP + TN}{TP + TN + FP + FN}, \end{aligned} \tag{2.1}$$

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 $\beta = 1$ формула (2.1) набуває виду F - міри (2.2). Метрика $g - mean$ (2.3), враховує правильні відповіді обох класів.

$$F1 = \left(1 + \beta^2\right) \cdot \frac{Precision \cdot Recall}{\left(\beta^2 \cdot Precision + Recall\right)}, \tag{2.2}$$

$$g_{mean} = \sqrt{\left(\left(\frac{TP}{TP + FN}\right) \cdot \left(\frac{TN}{TN + FP}\right)\right)}, \tag{2.3}$$

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

2.3 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 [26]. 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 (2.4):

$$n = 4 \times \delta^2 \times \left(z_{\alpha/2} / \omega \right)^2 + z_{\alpha/2}^2 / 2 \tag{2.4}$$

Поправочний коефіцієнт $z_{\alpha/2}^2$ вноситься у разі, що значення критерія z менше значення t-критертерію [26]. Виходячи з проведених експериментів, визначена довжина підпослідовності - 9 біт. Грунтуючись на описі статистичного тесту на підрахунок не перетинаючих шаблонів пакету тестів NIST [27], визначимо середньоквадратичне відхилення і математичне очікування для підпослідовностей довжиною 9 біт (2.5):

$$\mu = \frac{M - m + 1}{2^m} \tag{2.5}$$

$$\sigma^2 = M \times \left(\frac{1}{2^m} - \frac{2 \cdot m - 1}{2^{2m}} \right), \tag{2.6}$$

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 (2.4) 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. 2.1 наведено набір файлів проведення експериментів, для вибору математичного апарату класифікатора.

Таблиця 2.1 - 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 [30].

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 [34] 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.

У роботі [35] також 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. 2.2.

Таблиця 2.2 - 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

2.4 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.

3 МОДЕЛЬ ПСЕВДОВИПАДКОВИХ ПОСЛІДОВНОСТЕЙ З ВРАХУВАННЯМ ЇХ СТАТИСТИЧНИХ ХАРАКТЕРИСТИК

3.1 Алгоритми машинного навчання класифікації псевдовипадкових послідовностей

Алгоритм побудови дерева рішень. Дерево рішень - орієнтована конструкція, коренем (початковою точкою) є вершина, з зв'язками, що відходять вниз, гілками. Якщо гілка перестає ділитися, вона є стомом сформованого дерева. Принцип побудови дерева рішень для вирішення задач класифікації можна описати наступним чином (3.1):

$$L\left(Y, f\left(X\right)\right) \to \min, \tag{3.1}$$

де L – функція втрат, $f(X)$ – функція визначення класу набору даних, розмірності p $X = \left\{x_1, x_2, ..., x_p\right\}$. Функція втрат визначається як міра, наскільки правильно функція $f\left(X\right)$ визначає клас з множини X відповідно до множини Y.

Для задачі класифікації функція втрат визначається (3.2):

$$L\left(Y, f\left(X\right)\right) = I\left(Y \neq f\left(X\right)\right) = \begin{cases} 0, \textit{ якщо } Y = f(X); \\ 1, \textit{ інакше}. \end{cases} \tag{3.2}$$

Сформоване дерево рішень графічно представлене на рис. 3.1. Для розбиття вибірки даних необхідно вибрати ознаку за значенням якої здійснюватиметься розбиття. Правило вибору ознаки виражається наступним чином: атрибут повинен розділити множину вхідних даних так, щоб дочірні підмножини складалися з об’єктів, що містять мітки одного класу, кількість об'єктів, «домішок» (мітки інших класів) у підмножині мінімізовано. Інформаційний критерій Шеннона (3.3):

$$H = -\sum_{i=1}^{n} \frac{N_i}{N} \cdot \log\left(\frac{N_i}{N}\right) \tag{3.3}$$

Дерева поділяють дані на дві множини, значення ознаки яких не задовольняють і задовольняють умову розбиття у вузлі.

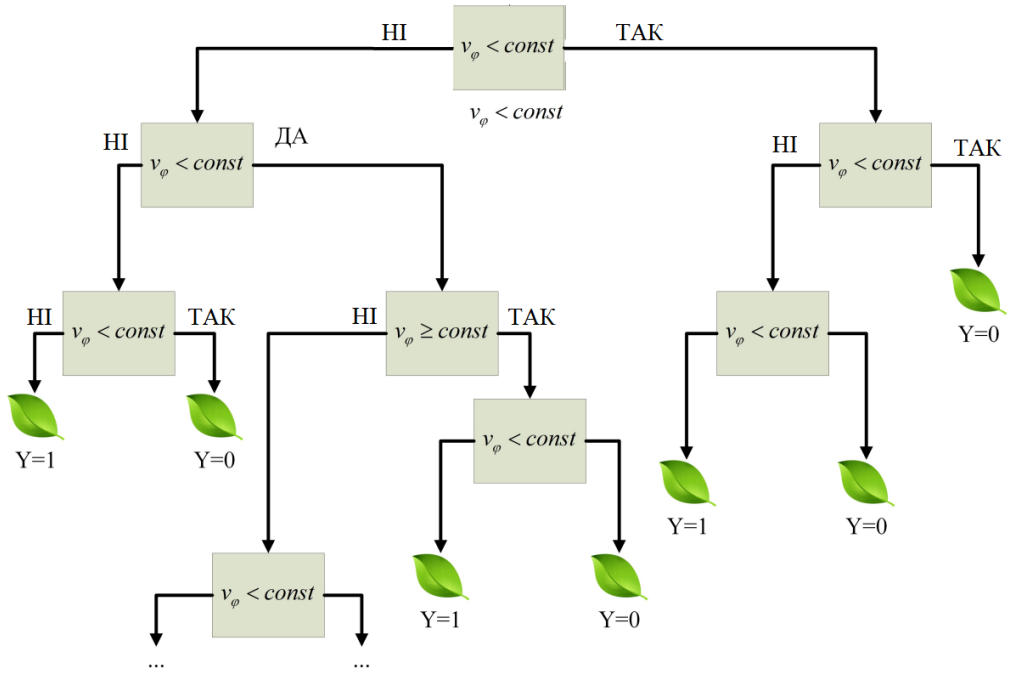


Рисунок 3.1 - Дерево рішень на основі ознак v_φ

Вершина дерева містить у собі весь ознаковий простір. Вибір певної точки розбиття відбувається на основі аналізу всіх можливих варіантів розбиття за ознаками за певним критерієм, в задачах класифікації використовуються статистичний критерій невизначеності та інформаційна ентропія Джині (3.4):

$$Q = \sum_{k \neq k`}^{K} p_k \cdot p`_k, \tag{3.4}$$

де p_k - часткою елементів класу k у вузлі і визначається виразом (3.5):

$$p`_k = \frac{1}{n} \cdot \sum_{i=1}^{n} I\left(y_i = k\right) \tag{3.5}$$

Критерій розбиття (вірність класифікації) - міра чистоти у вузлі, більше значення свідчить про наявність даних із різних класів. Після побудови дерева рішень на вхід подаються дані, що піддаються класифікації. Для визначення узагальненої помилки використовується інформація, яка входили в тренувальний набір, оскільки класифікатор покаже необгрунтовано високі результати. Дані які не брали участь у навчанні класифікатора, вводиться помилка узагальнення (3.6):

$$E_{вн} = \frac{1}{N} \cdot \sum_{i=1}^{N} I\left(y_i = f_{вн}\left(x_i\right)\right), \tag{3.6}$$

де f_{oob} – функція визначення класу даних, які не брали участь у навчанні класифікатора, y_i – істинний клас даних, I – функція втрат.

Дерева рішень можуть будуватись різними алгоритмами: CART, C4.5, ID3.Алгоритм ID3 представляє собою ітеративний дихотомайзер. Основна ідея алгоритму ID3 полягає у побудові дерева на основі перевірки кожного атрибуту в у вузлі дерева та жадібного пошуку за набором даних. Для розбиття можуть застосовуватись лише категоріальні атрибути. До недоліків алгоритму слід віднести необхідність передобробки даних та невисоку точність. Алгоритм C4.5 є покращенням ID3, розроблений покращена процедура відсікання гілок, дозволила знизити вплив надто деталізованих ознак та шумів у даних. Основними перевагами алгоритму є: можливість роботи з пропусками в даних; обробка атрибутів з різними вагами; покращена процедура відсікання гілок; обробка безперервних та дискретних значень атрибутів.

Алгоритм CART - використовує ітераційну парадигму, заснований на бінарному розбиття даних за атрибутами. Основна відмінність - висока точність у задачах класифікації, задачах прогнозування, можливість використання в

регресійному аналізі. Дані на вході алгоритму - незалежні вибірки випадкових величин $\left(X^1,...,X^p,Y\right)$, де: X^k - аналізовані дані; Y - категоріальна прогнозована змінна. Алгоритм CART здійснює розбиття вхідної множини на дві підмножини, щодо критерію однорідності, найкращим чином.

Алгоритм побудови випадкового лісу надзвичайно успішний у задачах регресії та класифікації, реалізує принцип навчання з учителем. Запропонований алгоритм поєднує декілька дерев рішень, побудованих на відібраних випадково ознаках, демонструє високу продуктивність. Випадковий ліс може використовуватися для задач класифікації, регресії. Випадкові ліси, з обчислювальної точки зору мають переваги: відносно швидкі як для передбачення так і для навчання; можуть вирішувати задачі класифікації та передбачення; невелика кількість настроюваних параметрів; можуть використовуватись у паралельній архітектурі обчислень; застосовуються для вирішення складних завдань; мають вбудовану оцінку помилки узагальнення. Випадковий ліс відноситься до методів побудови ансамблів, кожне дерево будується на основі змінних, які випадковим чином вибираються з набору даних.

Ансамбль дерев виражає функцію класифікації $f\left(x\right)$, функція визначається набором базових класифікаторів $h_1\left(x\right),...,h_J\left(x\right)$. Функція $f\left(x\right)$ реалізує підхід голосування, визначається клас, який набрав найбільшу кількість дерев (голосів). Враховуючи (3.2), функцію побудови дерева можна представити у виді (3.7):

$$f\left(x\right)=\arg\max_{y\in Y}\sum_{\psi=1}^{\Psi}I\left(\psi\left(x\right)\right), \tag{3.7}$$

де $\psi\left(x\right)$ - дерево на основі випадково обраних ознак *x*; *y* - клас з множини *Y*; $I\left(\psi\left(x\right)\right)$– цільова функція приросту даних, визначається індексом Джині або критеріями ентропії. Дерева $\psi\left(x\right)$ в ансамблі лісу засновані на рекурсивних розділяючих деревах.

Сформований класифікатор на основі лісу випадкового при класифікації даних застосовує схему голосування, при якій клас визначається на основі простої більшості голосів базових класифікакаторів (3.8):

$$Y = \begin{cases} 1, \ якщо \dfrac{1}{\Psi}\sum_{\psi=1}^{\Psi} y(\psi) > \dfrac{1}{2} \\ 0, \ інакше \end{cases} \tag{3.8}$$

Найпоширеніша помилка полягає в тому, що помилка випадкового лісу не може обчислюватися як середнє значення помилок всіх дерев, використовується середнє значення помилок кожного дерева, даний підхід дозволяє визначати помилку класифікації класу і будувати, для класифікатора, матрицю помилок шляхом порівняння значень і $f_{вн}(x_i)$ і y_i. Однією з переваг випадкового лісу є стійкість до незбалансованих наборів даних.

Для вирішення задач класифікації застосовується алгоритм *k*-найближчих сусідів, класифікує об'єкти на основі оцінки міток найближчих сусідів у просторі ознак. Параметр *k* - число сусідніх об'єктів у багатовимірному просторі ознак, які порівнюються з класифікуємим об'єктом. Графічна інтерпретація, з параметром *k* = 5, алгоритму *k* - найближчих сусідів представлена на рис. 3.2.

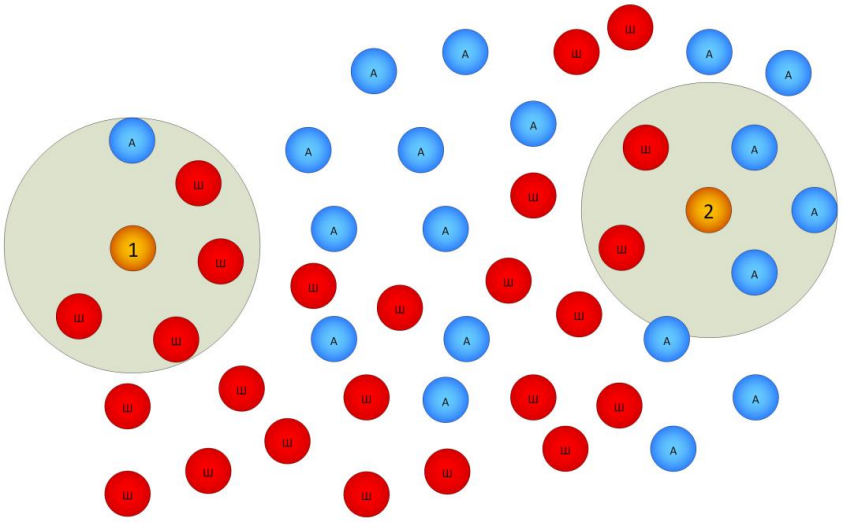


Рисунок 3.2 - Графічна інтерпретація алгоритму *k*-найближчих сусідів

Для об'єктів, з номерами 1 і 2 вибрано по п'ять найближчих сусідніх об'єктів. Для об'єкта 1 чотири об'єкти належать до класу шифрів, тому буде класифікований як шифр, для об'єкта 2 три найближчих сусіди, як архіви.

Метод опорних векторів виконує пошук гіперплощини (найбільш точної лінії), яка поділяє на два класи екземпляри даних, які розташовані до лінії поділу найближче. За формулою (3.9) обчислюється відстань -зазор.

$$t = \frac{m}{\|w\|}, \tag{3.9}$$

де m – відстань між межею і найближчими до неї екземплярами даних, виміряна вздовж вектора w; $\|w\|$ – нормоване значення вектора w. Можливий вибір для параметрів $t, m, \|w\|$ різних значень, найбільш часто задають значення m=1, тоді максимізація зазору відповідає мінімізації $\|w\|$, за умови відсутності екземплярів даних у зазорі, що призводить до задачі квадратичного оптимізації (3.10):

$$w^*, t^* = agr\min_{w,t} \frac{1}{2} \cdot \|w\|^2, \; f_i\left(w \cdot x_i - t\right) \geq 1, 1 \leq i \leq n \tag{3.10}$$

Графічна інтерпретація алгоритму опорних векторів представлена рис. 3.3.

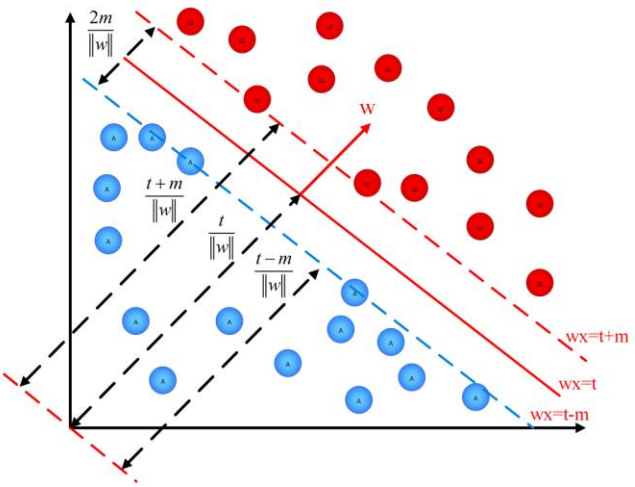


Рисунок 3.3 - Графічна інтерпретація методу опорних векторів

На рис. 3.3 наведено два опорні вектори, що проходять крізь екземпляри даних, що найближче знаходяться до в межі: три екземпляри (клас архівів) у нижній частині і два екземпляри (клас шифрів) у верхній частині рисунка.

Методи глибокого навчання застосовуються у задачах забезпечення безпеки інформаційної, відносяться до машинного навчання. Даний підхід заснований на використанні нейронних мережах для параметрів мережевого обладнання, класифікації об'єктів, поведінки процесів та користувачів, для ризиків виникнення інцидентів безпеки інформаційної, виявлення аномалій. До алгоритмів машинного навчання у задачах забезпечення безпеки інформаційної відносять: автоенкодери; згорткові нейронні мережі; глибокі мережі довіри; генеративно-змагальні мережі; навчання на основі безперервного зворотного зв'язку; рекурентні нейронні мереж.

Згорткові мережі - багатошарові ієрархічні мережі прямого поширення, в яких кожен шар на основі фільтрів (бази ядер) виконує множинні перетворення для отримання ознак, що володіють дискримінаційною характеристикою (певні області чи об'єкти). Операція згортки дозволяє виділити кольори, написи, контури елементи зображень. Вихід згорткових шарів пов'язаний з елементом нелінійним, дозволяє виявляти неявні ознаки даних. Нелінійність формує шаблони активації нейронів залежно від вхідних даних, дозволяє розрізняти і виявляти семантичну складову зображення. Процес навчання відбувається з використанням лгоритму зворотного розповсюдження помилки, дозволяє змінювати між нейронами ваги зв'язків на відповідну величину, що визначає напрямок та швидкість навчання, дозволяють досягти мінімальної помилки при розпізнаванні зображень.

Багатошарова ієрархічна структура мережі дозволяє витягувати з зображень слабопомітні, очевидні, непомітні ознаки. Графічна інтерпретація згорткових нейронних мереж наведена на 3.4. На теперішній час використовується концепція передачі навчання), навчена для будь-якої задачі мережа використовується для вирішення подібної задачі з донавчанням.

Автокодувальники - нейронні мережі прямого розповсюдження, моделюють вхідні дані на виході, у своїй архітектурі містять прихований шар, що виконує

процес виділення значущих ознак, дозволяють відновити вихідний сигнал на виході. Особливість – відповідність кількості нейронів на вході та виході мережі.

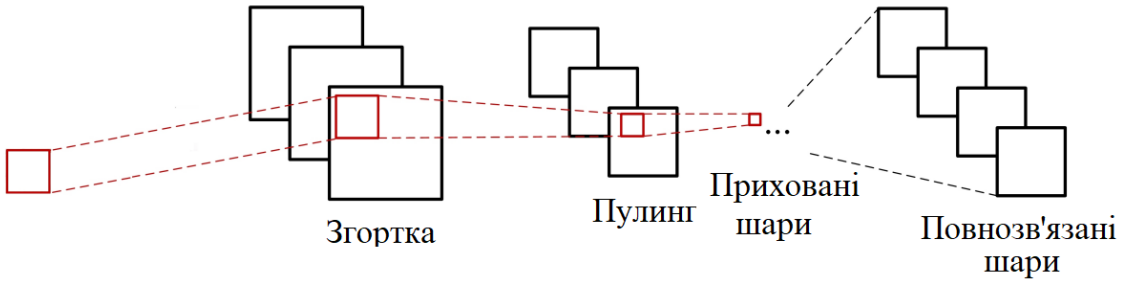


Рисунок 3.4 - Архітектура згорткових нейронних мереж

Автокодувальники складаються з двох частин: декодувальника та кодувальника. Декодувальник виконує зворотну процедуру, із прихованого шару отримуючи дані. Кодувальник виконує стиснення вихідної інформації і передачу на прихований шар. Прихований шар виділяє в аналізованих даних, значні ознаки, для цього необхідно виконання наступних вимог: дані, що подаються на вхід автоенкодера, повинні бути оптимізовані; рівність вихідного та вхідного шару; розмірність прихованого шару менше розмірності вхідного шару (рис.3.5).

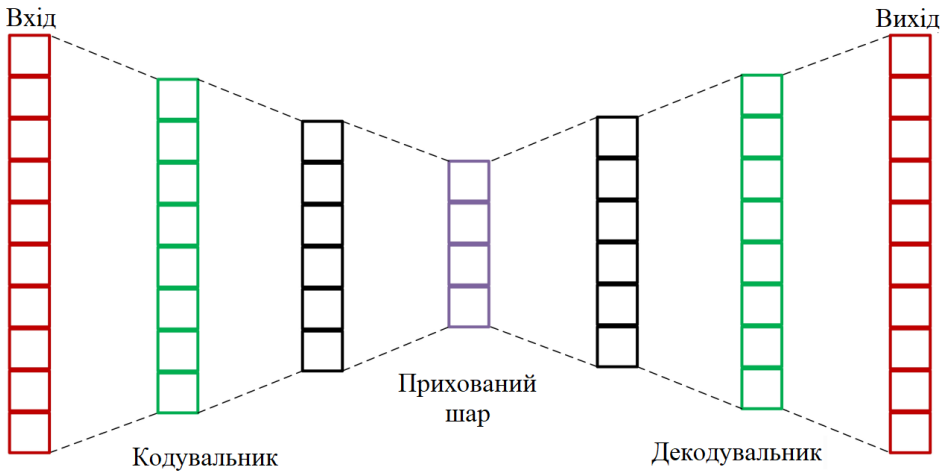


Рисунок 3.5 - Архітектура автокодувальників

Глибокі мережі довіри запропоновані для акустичного моделювання для розпізнавання мови, мають більш високу здатність за параметрами моделювання,

ніж Гауссівські моделі, мають досить ефективну процедуру навчання, яка поєднує в собі генеративне неконтрольоване навчання для виявлення ознак. Основним у мережах довіри є прихований шар, виконує пошук ознак, що визначають кореляцію вхідних даних. Графічна інтерпретація архітектури мереж довіри наведена на рис. 3.6. Мережі довіри - сукупність обмежених машин Больцмана. Сусідні шари з'єднані з використанням симетричної матриці ваг, при цьому відсутня повнозв'язна ієрархія.

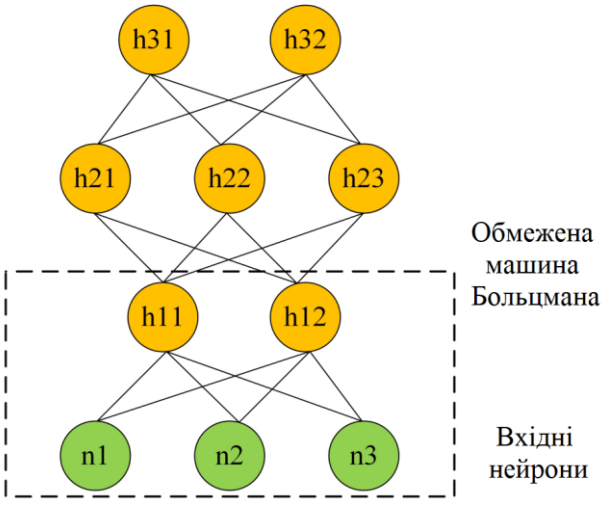


Рисунок 3.6 - Архітектура глибоких мереж довіри

Навчання мережі складається з наступних етапів: навчання мережі довіри пошарово з використанням алгоритму жадібного навчання; використання функції семплювання Гіббса для верхніх прихованих шарів; виконання функції налаштування ваг та навчання.

Відмінність рекурентних мереж полягає у послідовному використанні інформації. Рекурентні означає ітераційне виконання однієї функції обробки для кожного елемента вхідних даних, причому кожна настуна відповідь залежить від попередніх рішень. Рекурентні нейронні мережі можливо інтерпретувати як мережі, що мають пам'ять (враховують попередню інформацію). Графічна інтерпретація рекурентних нейронних мереж наведена на рис. 3.7. Процес визначення стану рекурентної нейронної мережі описується виразом 3.11:

$$\begin{cases} h_t = f_w\left(h_{t-1}, x_t\right) \\ h_t = \tanh\left(w_{hh} \cdot h_{t-1}, + w_{hx} x_t\right) \\ y_t = w_{hy} \cdot h_t \end{cases}, \tag{3.11}$$

де t - ітерація процесу навчання мережі; h – прихований шар; y - вихідні дані; x - вхідні дані; w– ваги зв‘зків мережі; $tanh$ – функція активації нейронів.

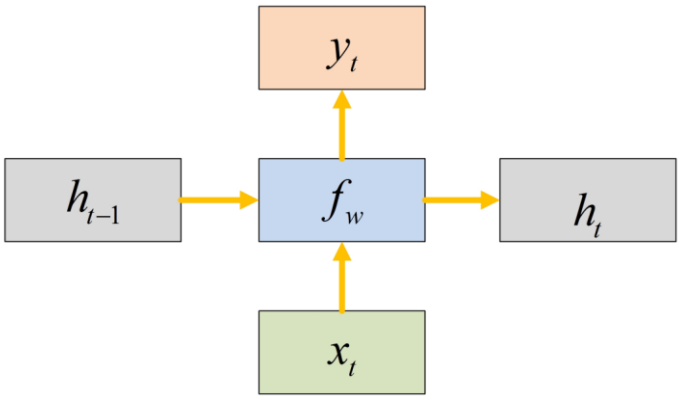


Рисунок 3.7 - Архітектура рекурентних нейронних мереж

В архітектурі змагальних мереж генеративна модель протиставляється дискримінаційній моделі – супротивнику, яка визначає, належить екземпляр даних моделі чи розподілу даних. Генеративну модель можна розглядати як аналог фальшивомонетників, які намагаються зробити валюту фальшиву і без виявлення використовувати, дискримінаційна модель аналогічна поліції, яка намагається виявити валюту фальшиву. Конкуренція у цій грі змушує команди вдосконалювати методи до тих пір, поки підробки стануть невідмінними від виробів справжніх.

Дискримінаційна модель мережі виконує класифікацію вхідних даних, аналізуючи псевдовипадкові послідовності і ознаки, дискримінаційний алгоритм передбачає, чи є вона послідовністю алгоритму стиснення даних або шифрування. У формальному виді задача полягає в обчисленні ймовірності $p(x|y)$, знаходженні ймовірності приналежності псевдовипадкові послідовності до

відповідного класу. Дискримінаційні моделі виконують зіставлення категорії з аналізованими даними. Генеративні моделі виконують протилежну функцію, замість класифікації даних формують екземпляри даних конкретного класу.

Дискримінаційні моделі встановлюють кореляційні ознаки між x і y, генеративні моделі здійснюють пошук закономірностей формування даних. Можна виділити особливості моделей: генеративні моделі дозволяють виділити особливості розподілу об'єктів; дискримінаційні моделі вивчають границю між класами об'єктів. Архітектура генеративно-змагальних мереж наведена на рис. 3.8.

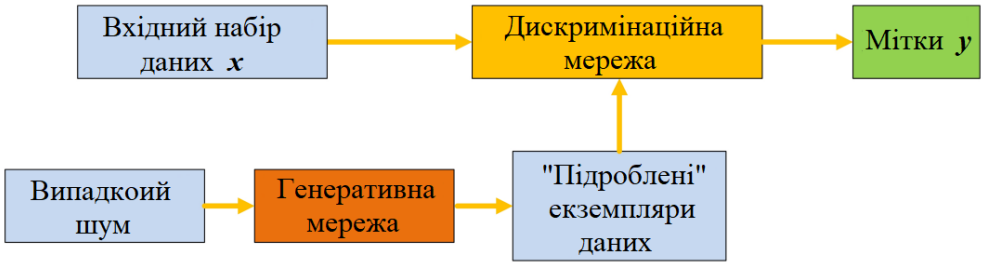


Рисунок 3.8 - Архітектура генеративно-змагальних мереж

3.2 Оцінка часової складності роботи алгоритмів машинного навчання

Для вибору відповідного алгоритму машинного навчання для побудови класифікатора псевдовипадкових послідовностей, як математичного апарату необхідно оцінити час, затрачений на безпосередньо класифікацію послідовностей та процедуру навчання.

Для оцінки складності алгоритму, функцією $f(n)$, використаємо вираз $O(g(n))$, де: $g(n)$ – формула, що виражає складність алгоритму; n - кількість даних, що подаються на вхід алгоритму. Дозволяє оцінити асимптотичний підхід, зміну часу виконання операцій в залежності від зміни об’єму вхідних даних. Формальний опис оцінки складності алгоритмів, визначається виразом 3.12.

$$\begin{cases} k \cdot f(n) \to O(g(n)), & якщо\ f(n) \to O(g(n)) і\ k = const, k > 0; \\ f(n) + h(n) \to O(g(n)) + p(n), & якщо\ f(n) \to O(g(n)) і\ h(n) \to O(p(n)); \\ f(n) \cdot h(n) \to O(g(n)) \cdot p(n), & якщо\ f(n) \to O(g(n)) і\ h(n) \to O(h(n)); \\ f(n) \to O(h(n)), & якщо\ f(n) \to O(g(n)) і\ g(n) \to O(h(n)); \\ f(n) \to O(n^k) & якщо\ f(n) \to n^k; \\ \log(nk) \to O(\log(n)), & якщо\ k > 0. \end{cases} \quad (3.12)$$

Перша властивість відноситься до коефіцієнтів, що до вхідних даних не належать, значення коефіцієнтів стають не значущими у разі устремління вхідних даних до нескінченності. Якщо часова складність алгоритму описується сумою різних функцій мчасової складності, нотація *О- велике* в даному випадку буде даною сумою. Правило множення часової складності алгоритму постулює принцип, як правило суми для мультиплікативної форми. Транзитивне правило означає, що алгоритми мають однакові значення нотації *О-велике,* якщо мають однакову часову складність. Поліноміальна залежність часової складності алгоритму призводить до тієї ж міри значення *О-велике* від вхідних даних. Винос коефіцієнта з-під логарифму означає залежності нотації *О-велике*, що і правило виносу коефіцієнта.

На підставі наведених властивостей, проведено аналіз часової складності алгоритмів машинного навчання, отриманні результати представлені в табл. 3.1. Для часової оцінки складності алгоритмів використані наступні позначення: Ψ − кількість класифікаторів в ансамблі; n − кількість вхідних послідовностей; m− кількість ознак у моделі псевдовипадкової послідовності; h − кількість нейронів; k − розмір кластера для алгоритму KNN; i − кількість ітерацій; q − кількість прихованих шарів; o − кількість виходів.

Алгоритм побудови мережі випадкового лісу починає формувати ансамбль класифікаторів - дерева рішень. У разі збільшення числа дерев до Ψ, збільшується часова складність. Пошук оптимального рішення полягає в пошуку оптимальних параметрів: мінімальної глибини, найменшої кількості дерев.

Таблиця 3.1 - Оцінка точності класифікації псевдовипадкових послідовностей на основі розподілу частот та байт послідовностей довжиною 9 біт

№ п/п	Алгоритм	Часова складність навчання	Часова складність класифікації
1	Decision Tree	$O\left(m \cdot n \cdot \log_2(n)\right)$	$O\left(\log_2(m)\right)$
2	Random Forest	$O\left(\Psi \cdot m \cdot n \cdot \log_2(n)\right)$	$O\left(\Psi \cdot m\right)$
3	k Nearest Neighbors	$O\left(m \cdot n^2\right)$	$O\left(m \cdot n\right)$
4	SVM	$O\left(n^2\right)$	$O\left(k \cdot m\right)$
5	Нейронні мережі	$O\left(m \cdot n \cdot h^q \cdot o \cdot i\right)$	$O\left(m \cdot n \cdot h^q \cdot o\right)$

На основі проведеного аналізу часової складності алгоритмів машинного навчання можна зробити наступні висновки: дерево рішень застосовується коли використовується великий набір даних; нейронні мережі мають досить трудомісткий процес навчання, найвищий ступінь складності; алгоритм SVM буде неефективним при використанні багаторозмірних ознакових просторів, на великих об’ємах даних, має найбільшу часову складність; ансамблевий метод побудови мережі випадкового лісу має більшу обчислювальну складність в порівнянні з дерева рішень.

Використання моделі псевдовипадкових послідовностей, яка враховує розподіл підпослідовностей довжиною 9 біт, рівномірний розподіл байт, точність класифікації послідовностей збільшилася при використанні розглянутих алгоритмів машинного навчання. Найвищу точність дали алгоритми *kNN*, проте час, що витрачається на класифікацію істотно перевищує аналогічний параметр у алгоритмів дерева рішень та випадкового лісу.

Для подолання недоліків необхідно розробити алгоритм класифікації псевдовипадкових послідовностей, що враховує ваги ознак, і що редукує ознаковий простір. Найбільш підходящим для побудови класифікатора

математичним апаратом є алгоритм побудови мережі випадкового лісу, має механізими визначення ваг ознак, досить високу точність класифікації псевдовипадкових послідовностей, мінімальний час формування класифікатора.

3.3 Модель псевдовипадкових послідовностей сформованих алгоритмами стиснення та шифрування інформації

Вибір ознак, що дозволять навчити класифікатор послідовностей і виконати процедуру класифікації із максимальною швидкістю і заданою точністю є задачею нетривіальною. Для класифікації відкритого та зашифрованого трафіку застосовувалися ентропійні підходи, що оцінюють ентропію блоків, потоків даних. Розглянуті рішення демонстрували високу точність класифікації відкритих та зашифрованих даних. Для класифікації відкритих і зашифрованих даних використовувалися отримані результати проходження послідовностями статистичних тестів з пакету NIST, при класифікації стислих та зашифрованих даних в якості просторових ознак використовувалися результати тесту приблизної ентропії, тесту кумулятивних сум, виконання блочного тесту. Для класифікації стислих і зашифрованих даних також використовувалися ентропійні підходи, проте високої точності не вдалося досягти.

Розглядалися різні підходи, що використовують обчислення ентропії потоку даних довжиною N-байт, враховучи, при цьому, розподіл частот байт що зустрічаються в послідовностях і статистичні ознаки: мінімальне і максимальне значення частот, середньоквадратичне відхилення, математичне очікування. Найбільш високої точності стиснутих даних класифікації 0,842 вдалося досягти при комбінуванні просторових ознак. Також, на теперішній час, існує досить велика кількість методів обходу сигнатурних підходів виявлення шкідливого програмного забезпечення та необхідності, заснованих на характеристиках бінарних послідовностей, застосування статистичних підходів.

Виходячи з аналізу поведених досліджень, існуючих підходів до класифікації псевдовипадкових послідовностей, сформованих алгоритмами стиснення та шифрування даних та враховуючи, також, дослідження в області машинного навчання можна висунути гіпотезу про наявність у стислих і зашифрованих даних статистичних особливостей та можливості виявлення цих особливостей методами машинного навчання.

Враховуючи, заголовну частину в структурі файлів, яку можна модифікувати для проведення маскування інформації, даний підхід може бути використаний зловмисниками для передачі конфіденційної інформації. Таким чином, необхідно видалення заголовків файлів, оскільки в них розміщені цифрові сигнали, що дозволяють, з високою точністю, класифікувати тип даних. При побудові моделі псевдовипадкових послідовностей відкинуті перші 10 кбайт даних файлів, щоб виключити вплив цифрових сигнатур та заголовків файлів на процедуру класифікації. На першому етапі побудови моделі ПВП проведені оцінки розподілу байт двох класів послідовностей, нормованих за середнім значенням частоти байта у вибірці даних, результати наведені на рис. 3.9.

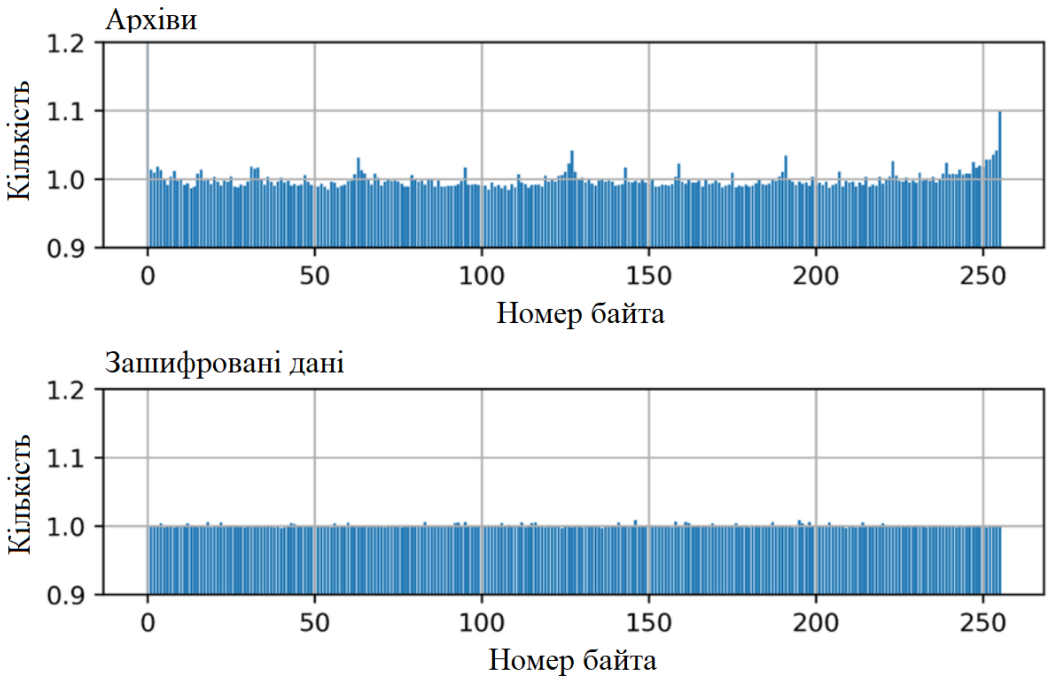


Рисунок 3.9 - Розподіл байт для стислих та зашифрованих послідовностей

Виходячи з рис. 3.9, про те, що розподіл байт стиснутих послідовностей більш нерівномірний ніж розподіл зашифрованих. Для оцінки отриманих розподілів частот байт у ПВП проведена оцінка зустрічаємості байт відповідності рівномірному розподілу згідно з критерієм Хі-квадрат (3.13):

$$\begin{cases} O_{rc} = \sum_{i=1}^{k} \dfrac{\left(x_i - m_i\right)^2}{m_i} \\ X^2 = \sum_{r=1}^{R} \sum_{c=1}^{C} \dfrac{\left(O_{rc} - E_{rc}\right)^2}{E_{rc}} \end{cases}, \tag{3.13}$$

де m_i - математичне очікування частоти появи байта $i \in \{0,...,255\} \in \{0, \ldots ,255\}$ в псевдовипадковій послідовності, x_i - значення частоти появи байта в послідовності, O_{rc}, E_{rc} - очікуване та спостережуване значення критерію на вибірці даних розміром r рядків та c стовпців.

Отримані значення дозволяють зробити висновок, у розглянутих послідовностях, про рівномірний характер розподілів байт і вважати їх псевдовипадковими. Ознаковий простір на основі розподілу байт (3.14):

$$V_{bytes} = \left\langle F\left(b_i\right)\right\rangle, i \in \left\langle 0,...,255\right\rangle, \tag{3.14}$$

де $F\left(b_i\right)$ - функція визначення частоти входження байта в псевдовипадкову послідовність. Результати експериментів щодо визначення точності класифікації зашифрованих послідовностей алгоритмами при використанні просторових ознак, що є значеннями розподілу частот появи байт наведені у виразі 3.15:

$$\begin{cases} Acc_{RF}\left(V_{bytes}\right) = 0.90 \\ Acc_{DT}\left(V_{bytes}\right) = 0.88 \\ Acc_{kNN}\left(V_{bytes}\right) = 0.89 \end{cases}, \tag{3.15}$$

де V_{bytes} – ознаковий простір, що визначається виразом 3.14; $Acc_{RF,DT,kNN}$ - частка правильних відповідей (метрики) при класифікації відповідним алгоритмом машинного навчання псевдовипадкових послідовностей.

Отримані результати дозволяють побудувати класифікатор стиснених і зашифрованих послідовностей, точність класифікації, при цьому, недостатньо висока. Розподіл байт для послідовностей, що мають високу ентропію, для стиснених і зашифрованих даних, підпорядковується закону рівномірного розподілення. Для побудови моделі псевдовипадкових послідовностей недостатньо лише розподіл байт, оскільки вони сягають до рівномірного розподілу. Необхідно знайти новий ймовірнісний простір ознак, який дозволить створити більш точну модель псевдовипадкових послідовностей, були проведені обчислення середніх значень ентропії Шеннона (3.16):

$$H = \sum_{i=0}^{255} p_i \cdot \log_2 p_i, \tag{3.16}$$

де p_i – ймовірність появи байта $i \in \{0,...,255\}$ в послідовності що аналізується. Отримані значення статистичних параметрів розподілів ентропії байт та середні значення ентропії в послідовностях, що аналізуються, результати наведені в табл. 3.2, отримані відповідно до виразу 3.17:

$$\begin{cases} H_{mean} = \dfrac{\sum_0^{255} H(b_i)}{256} \\ H_{sko} = \sqrt{\dfrac{\sum_0^{255} \left(H(b_i) - H_{mean}\right)}{256}} \\ H_{median} = X_{median} + i_{median} \cdot \dfrac{\dfrac{\sum_0^{255} H(b_i)}{2} - sum_{median-1}}{H(b_i)} \end{cases}, \tag{3.17}$$

де $H(b_i)$ - ентропія b_i-го байта; H_{mean} – середнє значення ентропії; H_{sko}– середнє квадратичне відхилення ентропії; H_{median} – медіанне значення ентропії.

Таблиця 3.2 - Статистичні ознаки розподілу значень ентропії байт

Ознака	Архіви	Шифри
H_{mean}	5.543	5.545
H_{sko}	0.001	0.0002
H_{median}	5.544	5.545

Для побудови моделі псевдовипадкових послідовностей використано статистичні тести NIST. Елементами моделі - значення p-value, отримані в результаті проходження статистичних тестів. Модель псевдовипадкових послідовностей на основі значень p-value тестів NIST є усередненими значеннями p-value, отримані в результаті статистичних тестів. Для оцінки ознакового простору проведені експерименти, результати представлені у табл. 3.3. При використанні моделі псевдовипадкових послідовностей на основі тестів NIST точність класифікації порівняно з моделлю розподілу байт значно погіршилась, тести NIST спрямовані на виявлення у аналізованих послідовностях закономірностей, перевірку на випадковість виникнення символів.

Таблиця 3.3 - Оцінка точності класифікації послідовностей на основі NIST

№ п/п	Алгоритм	Accuracy
1	Random Forest	0.644
2	Decision Tree	0.576
3	k-Nearest Neighbors	0.685

На підставі даного тесту висунуто припущення про можливість побудови моделі псевдовипадкових послідовностей на основі частот, обмеженої довжини, зустрічальності підпослідовностей. Для виявлення статистичних особливостей в

аналізованих псевдовипадкових послідовностях, проведені експерименти з підрахунку частот входження підпослідовностей довжини N = [4, . . . 11] біт. Підрахунок частот входження підпослідовностей виконаний відповідно до 3.18:

$$f_j = F(j) = \frac{n(j)}{M - N(j) + 1}, j \in \left\{0,...,2^N\right\}, \quad (3.18)$$

де f_j - частота входження підпослідовності j в аналізовану послідовність, $n(j)$ - кількість входжень підпослідовності j в аналізовану псевдовипадкову послідовність, M – довжина аналізованої послідовності в бітах, $N(j)$ – довжина підпослідовності в бітах.

Даний підхід, на відміну від статистичних тестів, де використовуються неперіодичні шаблони певної довжини, дозволить визначити їх дискримінаційну здатність та перевірити всі можливі підпослідовності. Модель послідовностей - вектор статистичних характеристик, які обчислюються виразом 3.19:

$$V_{Sub} = \left(f_j,..., f_{2^N} \right) \quad (3.19)$$

Для визначення оптимальної довжини підпослідовності даних проведені експерименти, вхідна вибірка була перетворена на вісім наборів даних $V = V_4,...,V_{11}$, містять вектори псевдовипадкових послідовностей, які визначаються виразом 3.18, і складаються зі значення частот підпослідовностей довжиною 4-11 біт. Отримані ознакові простори подавалися на вхід алгоритму випадкового лісу для визначення точності класифікації псевдовипадкових послідовностей, результати наведені на рис. 3.10.

Найбільш раціональним значенням довжини підпослідовностей, залежно від часу і точності класифікації, є значення підпослідовності - 9 біт. Для перевірки запропонованої моделі на адекватність проведені експерименти з класифікації псевдовипадкових послідоввностей алгоритмами машинного навчання, отримані

результати наведені в табл. 3.4. Отримані результати свідчать про зниження точності класифікації алгоритмами дерева рішень та випадкового лісу при використанні моделі на основі врахування частот підпослідовностей довжиною в 9 біт, пточність класифікації алгоритмом kNN збільшилася.

Таблиця 3.4 - Оцінка точності класифікації алгоритмами машинного навчання при використанні моделі послідовностей довжиною 9 біт

№ п/п	Алгоритм	Accuracy
1	Random Forest	0.859
2	Decision Tree	0.858
3	k-Nearest Neighbors	0.902

Отриманий результат пояснюється тим, що модель містить 512 значень замість 256 при врахувані розподілу байт.

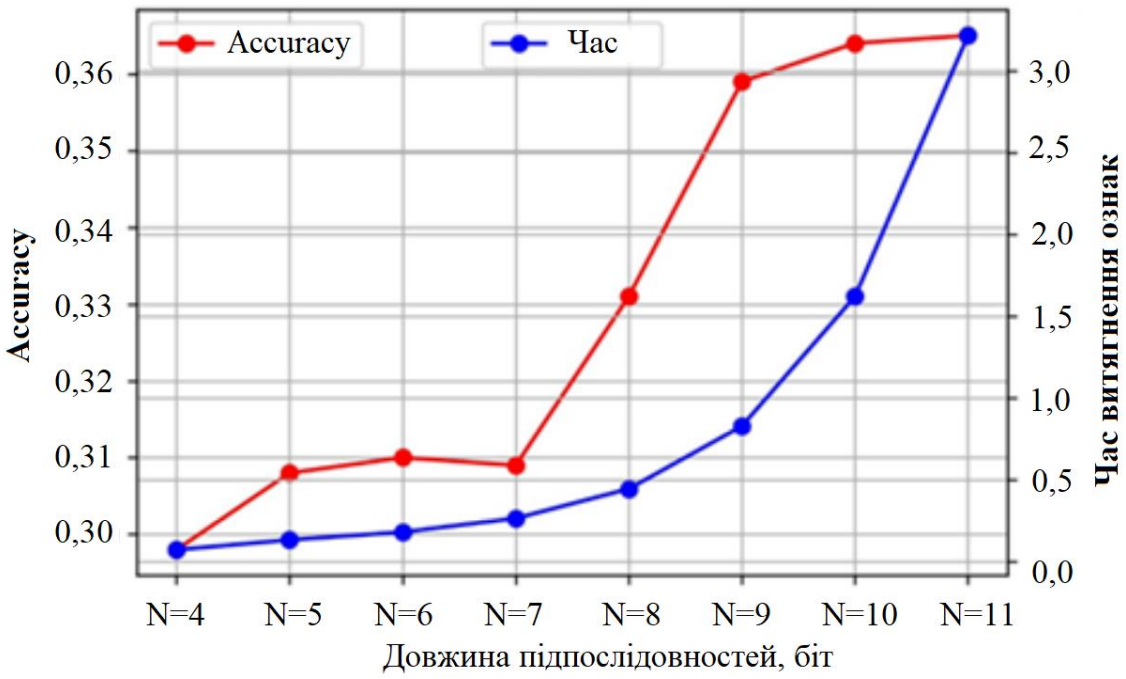


Рисунок 3.10 - Залежність точності класифікації псевдовипадкових послідовностей на основі алгоритмому випадкового лісу

Оскільки частоти підпослідовностей довжини 9 біт і розподіл байт мають різні імовірнісні простори і розподіли, то раціональним кроком у проведенні досліджень - побудова синтезованої моделі псевдовипадкових послідовностей. До моделі додані статистичні ознаки розподілу байт: середнє значення (B_{mean}), середньоквадратичне відхилення (B_{sko}), мінімальне ($b_{\min}$) та максимальні ($b_{\max}$) значення кількості байт у послідовності, що визначаються відповідно до 3.20:

$$\begin{cases} B_{mean} = \dfrac{\sum_{i=0}^{255} n(b_i)}{256} \\ B_{sko} = \sqrt{\dfrac{\sum_{i=0}^{255} \left(n(b_i) - B_{mean}\right)^2}{256}}, \\ b_{\min} = Min\left(n(b_i)\right) \\ b_{\max} = Max\left(n(b_i)\right) \end{cases} \tag{3.20}$$

де ($n(b_i)$) – кількість появи i - байта в псевдовипадковій послідовності аналізованій, яка піддається аналізу.

Модель псевдовипадкової послідовності - вектор статистичних характеристик, що обчислюється виразом 3.21:

$$V_{stat} = \langle v_1,...,v_\varphi \rangle = \langle f_j,..., f_{2^N}, b_0,...,b_{255}, B_{mean}, B_{sko}, b_{\min}, b_{\max} \rangle \tag{3.21}$$

Для виконання вимоги оперативності запропонованого методу захисту від витоку інформації на основі стиснених та зашифрованих даних розроблено алгоритм редукування простору ознак. Застосування запропонованого алгоритму дозволить зменшити вхідний розмір моделі псевдовипадковх послідовностей в сім разів, підвищивши точність класифікації послідовностей за рахунок пошуку дискримінуючих статистичних ознак.

3.4 Висновки

Запропонована модель псевдовипадкових послідовностей, відрізняється від відомих, на теперішній час, з урахуванням розподілу байт та частот бітових підпослідовностей довжиною в дев’ять біт.

Обґрунтовано вибір довжини підпослідовностей, довжиною в дев’ять біт, проведено аналіз найчастіше використовуваних просторів ознак при класифікації псевдовипадкових послідовностей. Сформовано вибірку стислих і зашифрованих даних, які в предметній області досліджень найчастіше зустрічаються, обґрунтовано розмір вибірки файлів. Запропоновано оцінку точності класифікації псевдовипадкових послідовностей алгоритмів машинного навчання, для використання у класифікаторі псевдовипадкових послідовностей, обґрунтовано вибір алгоритму побудови випадкового лісу.

Для підвищення оцінки точності класифікаторів запропонованоо алгоритм редукування простору ознак моделі псевдовипадкових послідовностей, що дозволяє підвищити точність класифікації послідовностей, знизити розмірність ознакового простору, на осові вибору найбільш значущих ознак.

4 МЕТОД КЛАСИФІКАЦІЇ ПСЕВДОВИПАДКОВИХ ПОСЛІДОВНОСТЕЙ СФОРМОВАНИХ АЛГОРИТМАМИ СТИСНЕННЯ ТА ШИФРУВАННЯ ІНФОРМАЦІЇ

4.1 Метод класифікації стиснених та зашифрованих псевдовипадкових послідовностей

Запропоновано метод, на основі класифікації стислих та зашифрованих даних захисту від витоку інформації та спосіб реалізації. Метод класифікації стислих і зашифрованих даних дозволяє підвищити точність класифікації при використанні моделі псевдовипадкових послідовностей за рахунок застосування комплексу алгоритмів з навчання класифікатора, виявлення найбільш значущих дискримінуючих ознак.

Вихідними даними для побудови класифікатора псевдовипадкових послідовностекй - множина послідовностей стислих і зашифрованих (два класи).

Метод класифікації враховує дискримінуючу здатність послідовностей статистичних ознак, складається з наступних кроків:

1. Відкидання перших десять кілобайт файла та підрахунок ентропії.

2. Визначення порогового значення ентропії файла;

3. Визначення режиму роботи: швидка, шляхом перевірки випадково вибраного вікна розміром 500 кбайт, повна перевірка файлу скануючим вікном розміром 500 кбайт.

4. Обчислення значень ознак згідно з моделлю ПВП.

5. Визначення типу файлу навченим класифікатором на основі градієнтного бустингу.

6. Якщо визначено клас послідовностей, то продовжити роботу алгоритму, інакше закінчити.

7. Обчислення значень ознак згідно з моделлю ПВП.

8. Передача отриманих значень ознак на навчений класифікатор.

9. Ітераційний рух вузлами дерев випадкового лісу.

10. Визначення досягнення термінального вузла.

11. Визначення класу псевдовипадкової послідовності.

Алгоритм класифікації псевдовипадкових чисел представлений на рис. 4.1.

На першому етапі роботи алгоритму здійснюється розбиття вхідного набору даних, який містить у собі стислі та зашифровані послідовності, на навчаючу і тестову підвиборки. В навчальну вибірку відноситься 80% виідних послідовностей, решта 20% відносять до тестової вибірки. Для забезпечення незалежності методу захисту від витоку інформації виконується відкидання перших десять кілобайт файлу, що аналізується та підрахунок ентропії.

На другому етапі роботи алгоритму відбувається визначення порогового значення ентропії. Для зашифрованих даних встановлено поріг 6,5. При перевищенні значення порогу файл, вважається підозрілим інакше - легітимний.

На третьому етапі, відбувається формування ознакового простору та обчислення значень статистичних ознак згідно з виразом 3.21. Ознаковий простір складається з 512 значень частот входження підпослідовностей довжиною дев’ять біт $\langle f_j \rangle$, $j \in [0,...,511], 256$ частот входження байт $\langle b_i \rangle$, $i \in [0,...,255]$ та чотири статистичних характеристик $\langle B_{mean}, B_{sko}, b_{\min}, b_{\max} \rangle$. Отриманий ознаковий простір складається з 772 ознак, використання класифікатора, що враховує отримані значення, призведе до збільшення часу класифікації і неможливості застосування алгоритму, в режимі реального часу. Для усунення обмеження пропонується використовувати метод полегшеного градієнтного бустингу.

На третьому етапі відбувається формування редукованої множини ознакового простору. Розмірності простору здійснюється за рахунок випадково обраних ознак, дають мінімальне зменшення градієнта та використання ознак, що дають максимальне зниження градієнта функції втрат. Вибір ознак з мінімальним градієнтом дозволяє зберегти точність класифікації, знизити ознаковий простір.

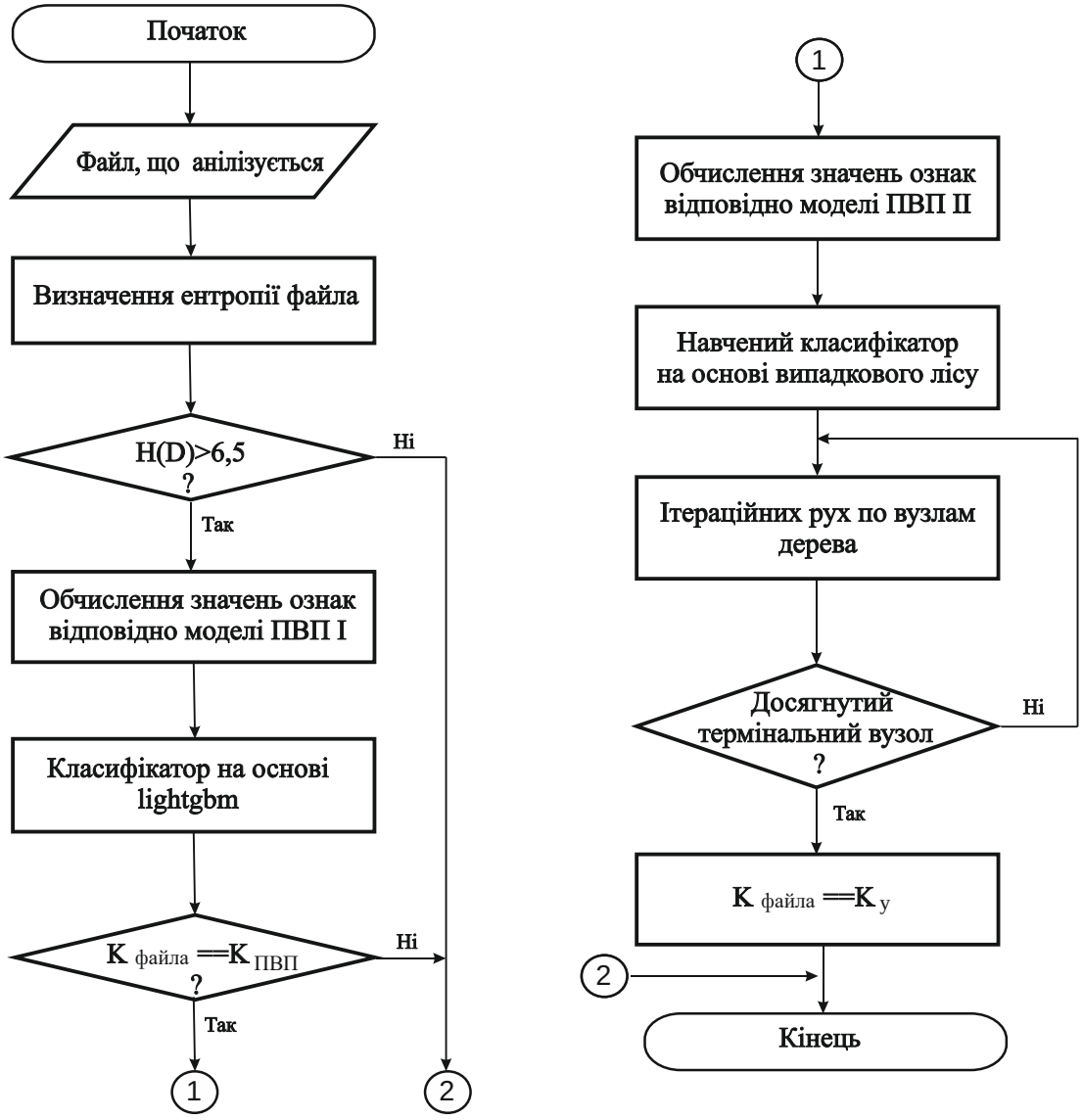


Рисунок 4.1 - Алгоритм класифікації стислих та зашифрованих даних

На четвертому та п'ятому етапах виконується розподіл підозрілих об'єктів, що містять у собі стислі чи зашифровані дані. У разі легітимного файлу робота алгоритму завершується, при виявленні підозрілого об'єкта виконується перехід на шостий етап. На даному етапі здійснюється вибір режиму проведення аналізу даних: вибірковий, дозволяє аналізувати випадково обраний фрагмент розміром 500 кілобайт послідовний, полягає в проходженні всього файлу скануючим вікном розміром 500 кілобайт.

На шостому етапі здійснюється обчислення ознак, редукованих за допомогою навченого класифікатора на основі випадкового лісу. Редукування

простору ознак виконується на основі обчислення локальних ваг. Критерієм розбиття визначається виразом 4.1.

$$\begin{cases} IG\left(D_p,\upsilon\right)=IG\left(D_p\right)-\dfrac{N_{left}}{N_p}\cdot I\left(D_{left}\right)-\dfrac{N_{right}}{N_p}\cdot I\left(D_{right}\right), \\ IG\to\max \end{cases} \tag{4.1}$$

де $IG\left(D_p\right)$– приріст інформації після розподілу батьківського вузла, $I\left(D_p\right)$, $I\left(D_{left}\right), I\left(D_{right}\right)$– значення міри неоднорідності в батьківському вузлі, правому та лівому нащадках відповідно, $\upsilon\in V$ – ознака, за якою відбувається розбиття даних. Помилка класифікації, міра неоднорідності Джині визначаються (4.2):

$$\begin{cases} Gini=1-\sum\limits_{y=1} p_y^2 \\ Entropy=-\sum\limits_{y=1} p_y\cdot\log_2 p_y \\ Error=1-\max\left(p_y\right) \end{cases}, \tag{4.2}$$

де p_y- ймовірність появи псевдовипадкової послідовності класу y в аналізованому вузлі класифікатора. На підставі отриманих значень визначаються локальні ваги ознак простору $w_{\upsilon_\varphi}^{\psi}$ для класифікаторів ψ відповідно до виразу 4.3.

$$w_{\upsilon_\varphi}^{\psi}=F\left(n_{\upsilon_\varphi}^{\psi}\right), \tag{4.3}$$

де $n_{\upsilon_\varphi}^{\psi}$ – порядковий номер ознаки υ_φ в дереві ψ відповідно до його дискримінуючої здатності; F – функція визначення ваги ознак у дереві ψ.

Локальні ваги ознак визначаються за різними мірами неоднорідності: помилка класифікації, міра неоднорідності Джині, ентропія. Враховуються, при цьому, різні ваги критеріїв у проміжку $\left[0{,}1\ldots 0{,}9\right]$, має виконуватися умова 4.4.

$$\begin{cases} \alpha + \beta + \gamma = 1 \\ 0 \le \alpha, \beta, \gamma \le 1 \end{cases} \tag{4.4}$$

Далі здійснюється визначення глобальних ваг відповідно до виразу 4.5:

$$\begin{cases} W_{\upsilon_\varphi} = F\left(\sum\limits_{\psi=1}^{\Psi} w_{\upsilon_\varphi}^{\psi} \right) \\ W_{\upsilon_\varphi}^{Global} = \min\limits_{\alpha,\beta,\gamma \in R} F\left(\upsilon_\varphi, \psi\right) = \alpha \cdot Entropy\left(\psi, w_{\upsilon_\varphi}^{\psi}\right) + \\ + \beta \cdot Gini\left(\psi, w_{\upsilon_\varphi}^{\psi}\right) + \gamma \cdot Error\left(\psi, w_{\upsilon_\varphi}^{\psi}\right) \end{cases} \tag{4.5}$$

де $w_{\upsilon_\varphi}^{\psi}$ – локальна вага ознаки υ_φ в дереві ψ; F – функція визначення глобальної ваги ознаки на основі більшості; Ψ – кількість дерев в ансамблі випадкового лісу.

Після пошуку найзначніших ознак, які мають максимальну дискримінуючу здатність, виконується сортування ознак починаючи з максимального значення за їх значимістю, здійснюється налаштування параметрів класифікатора: максимальна кількість класифікаторів в ансамблі; кількість ознак моделі псевдовипадкових послідовностей, що беруть участь у формуванні класифікатора; максимальна глибина класифікатора.

Для побудови кінцевого варіанта класифікатора використовуються ознаки, з найбільшую дискримінуючою здатністю і певними параметри, що дозволить скоротити час виконання класифікації.

На восьмому і дев'ятому етапах відбувається ітераційний рух по вузлах сформованих дерев, до моменту подання в термінальний вузол кожного дерева випадкового лісу. На десятому етапі, визначений клас надається аналізованому файлу, що аналізується. У разі присвоєння інформації мітки (стиснених) зашифрованих даних файл поміщається в карантин і відбувається генерація події безпеки даних відповідно до прийнятої політики в організації.

Алгоритм отримання простору ознак на основі моделі псевдовипадкових послідовностей. Оскільки сформований класифікатор є деревом рішень, то для класифікації псевдовипадкових послідовностей необхідно виконати дихотомічне проходження по вузлам сформованого класифікатора, для цього необхідно здійснити обчислення ознак псевдовипадкової послідовності згідно з алгоритмом, представлений псевдокодом на рис. 4.2.

Data: $P,|P| = Q,\ S : |S| = 512,\ B : |B| = 256$;
Result: V_{stat}

$V_{stat} \leftarrow <>$;
for $p \in P$ **do**
  for $s \in S$ **do**
    $n_s \leftarrow \mathbf{Count}(p,s)$;
    $f_{p,s} \leftarrow \dfrac{n_s}{M_p - N_s + 1}$;
    $f_{p,s} \leftarrow ln(f_{p,s})$;
    $V_{stat} \leftarrow f_{p,s}$
  for $b \in B$ **do**
    $n_b \leftarrow \mathbf{Count}(b,s)$;
    $bytes_p \leftarrow < b,n_b >$;
    $V_{stat} \leftarrow bytes_p$
  $V_{stat} \leftarrow \mathbf{Mean}(bytes_p)$;
  $V_{stat} \leftarrow \mathbf{SKO}(max_b,min_b)$;
  $V_{stat} \leftarrow \mathbf{Min}(bytes_p)$;
  $V_{stat} \leftarrow \mathbf{Max}(bytes_p)$;
return V_{stat};

Рисунок 4.2 - Алгоритм витягнення ознак із псевдовипадкових послідовностей

Вхідна послідовність $p \in P$ потужністю Q перетворюється на бінарний код. Далі здійснюється підрахунок зустрічальності підпослідовностей s довжиною дев'ять біт – значення n_s і визначення частот зустрічальності підпослідовностей $f_{p.s}$. Для кожного байта b визначається кількість входжень в послідовності.

Підрахунок числа входжень відбувається ковзним вікном без повторень. Для отриманого розподілу байт визначаються статистичні характеристики: середньоквадратичного відхилення, математичне очікування, максимальне та мінімального значення частот входження байт. Результат роботи алгоритму формується в кортеж.

4.2 Реалізація методу класифікації стиснених та зашифрованих псевдовипадкових послідовностей

Підсистеми статистичного аналізу інформації на основі методу класифікації псевдовипадкових послідовностей реалізована з використанням мови програмування Python, реалізує метод захисту від витоку інформації на основі розподілу стислих та зашифрованих даних, може бути використаний для виявлення мережевих атак на мережі передачі даних, в існуючі засоби запобігання та виявлення витоку інформації, а також в програмні продукти, що реалізують сервіси електронної пошти.

Підхід до раннього виявлення деструктивних впливів Botnet на мережу базується на формуванні мережі моніторингу комп'ютерних атак, включає канали зв'язку з сервером мережі управління системою захисту, мережні датчики, встановлені в сегментах корпоративної мережі. Проводиться моніторинг, у режимі реального часу, деструктивних впливів на мережу, формується база даних про параметри деструктивних впливів на вузли мережі. Під час моніторингу деструктивних впливів (ДВ) вимірюють значення параметрів впливів на об'єкти мережевої інфраструктури: кількість вузлів, що беруть участь у ДВ, тривалість ДВ, час отримання команд на початок ДВ. Після виявлення факту ДВ на підставі отриманих статистичних даних прогнозують параметри мережі, що функціонує в умовах ДВ, параметри ДВ; на підставі отриманих ознак простору ідентифікують зловмисника. На підставі від ДВ спрогнозованого збитку, визначають перелік

способів та варіантів протидії ДВ, порядок їх використання. На підставі проведеного аналізу параметрів впливу фіксують закінчення деструктивного впливу Botnet; після закінчення деструктивного впливу порівнюють фактичні значення параметрів ДВ і розраховані величини параметрів шкоди з наявними в базі даних, при перевищенні значень даних вносять зміни в базу даних. Якщо не відповідають заданим значенням, значення параметрів впливів Botnet, уточнюють величини параметрів. Оцінюють ефективність застосовуваних способів і варіантів протидії, якщо значення параметрів деструктивного впливу відповідають заданим. Якщо значення параметрів задіяної шкоди відповідають значенням у базі даних, продовжують моніторинг; якщо значення задіяної шкоди не відповідають значенням в базі даних, проводиться уточнення значення параметрів.

На основі отриманих статистичних даних визначають шкідливе програмне забезпечення, з використанням яких зловмисник заражає ПК і команди, що надсилаються центром управління Botnet, зараженим персональним комп'ютерам. Визначають, також, причини зараження персональних комп'ютерів шкідливим програмним забезпеченням Botnet: використовувана операційна система, браузер, найчастіше відвідувані сайти, налаштування засобів захисту або їх відсутність.

Налаштовують віртуальну машину-пісочницю (персональний комп'ютер), щоб підвищити ймовірність зараження шкідливим програмним забезпеченням Botnet, налаштовують засоби захисту. На фазі встановлення з'єднання, у разі початку ДВ, вузли Botnet одержують команди керування в зашифрованому виді.

З використанням запропонованого модуля статистичного аналізу інформації визначається зараження ПК шкідливими програмами Botnet, проводиться аналіз вхідного потоку даних на наявність команд керування Botnet. Отримані дані конвертуються в бінарний код, далі формується вектор статистичних характеристик згідно запропонованого алгоритму класифікації псевдовипадкових послідовностей. Якщо виявлено, що послідовності, сформовані криптоалгоритмами в даних немає, то виконується переадресація даних у центр очищення повідомлень, де після знищення (очищення) інформації, продовжується

обробка і прийом вхідного потоку даних. Якщо виявлено що послідовності, сформовані криптоалгоритмами, дані перенаправляються в центр дешифрування повідомлень. На підставі проведеного аналізу прийнятих повідомлень формується рішення про надходження команди від Botnet, якщо команди від Botnet на захищаємий вузол, не надійшло, то продовжується моніторинг. У випадку виявлення команд від Botnet визначається IP-адрес керуючих атакою Botnet вузлів та вид деструктивного впливу.

Алгоритм раннього виявлення атак Botnet на мережу наведений на рис. 4.3.

Результатом модуля статистичного аналізу - зниження часу прийняття рішення про протидію деструктивному впливу на мережу передачі даних з боку Botnet, за рахунок ідентифікації початку впливу на етапі підготовки Botnet до деструктивного впливу. Проблема вирішується за рахунок раннього виявлення деструктивних впливів Botnet на мережу передачі даних, за рахунок аналізу потоку даних що надходить на попередньо заражений ПК, визначають мету і вид впливів Botnet, на підставі проведеного аналізу вживають відповідних заходів щодо завчасної активації засобів протидії деструктивному впливу (рис. 4.4).

Сучасні засоби запобігання та виявлення витокам інформації застосовують різні методи проведення аналізу потоку даних. До основних відносяться контекстні та контентні методи. Контентні методи застосовують пошук регулярних виразів, цифрових сигнатур, шаблонів та аналіз тексту. Контекстні методи базуються на проведення аналізу службової інформації: номер порту одержувача та джерела даних, ip-адреса, розмір даних, величина міжпакетних інтервалів, протокол передачі, наявність прапорів. Наведені методи не здатні виявити витік даних у стислому та зашифрованому виді, а додавання цифрових сигнатур дозволяє маскувати зашифровані дані під стислі простим способом.

На теперішній час в області безпеки інформації знайшли широке використання поведінкові методи аналізу потоку даних та алгоритми машинного навчання. Однією з основних складностей, в даній ситуації, виступає побудоаи моделей даних, обробка та пошук ознакового простору, що дозволяють

алгоритмам машинного навчання з високою точністю проводити класифікаціїю даних, об'єкти та дії різних класів.

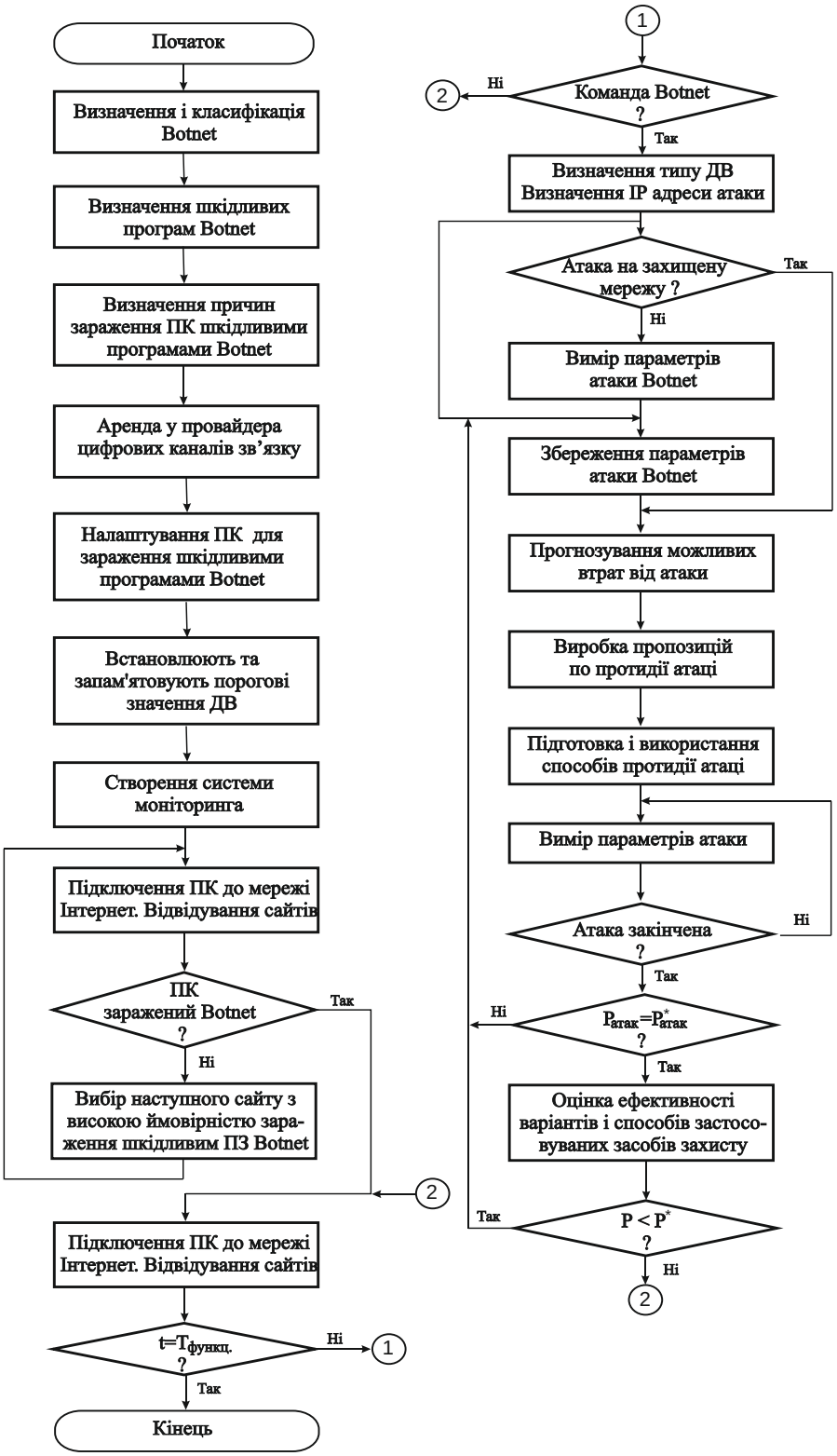


Рисунок 4.3 - Алгоритм раннього виявлення деструктивних впливів Botnet на мережу передачі даних

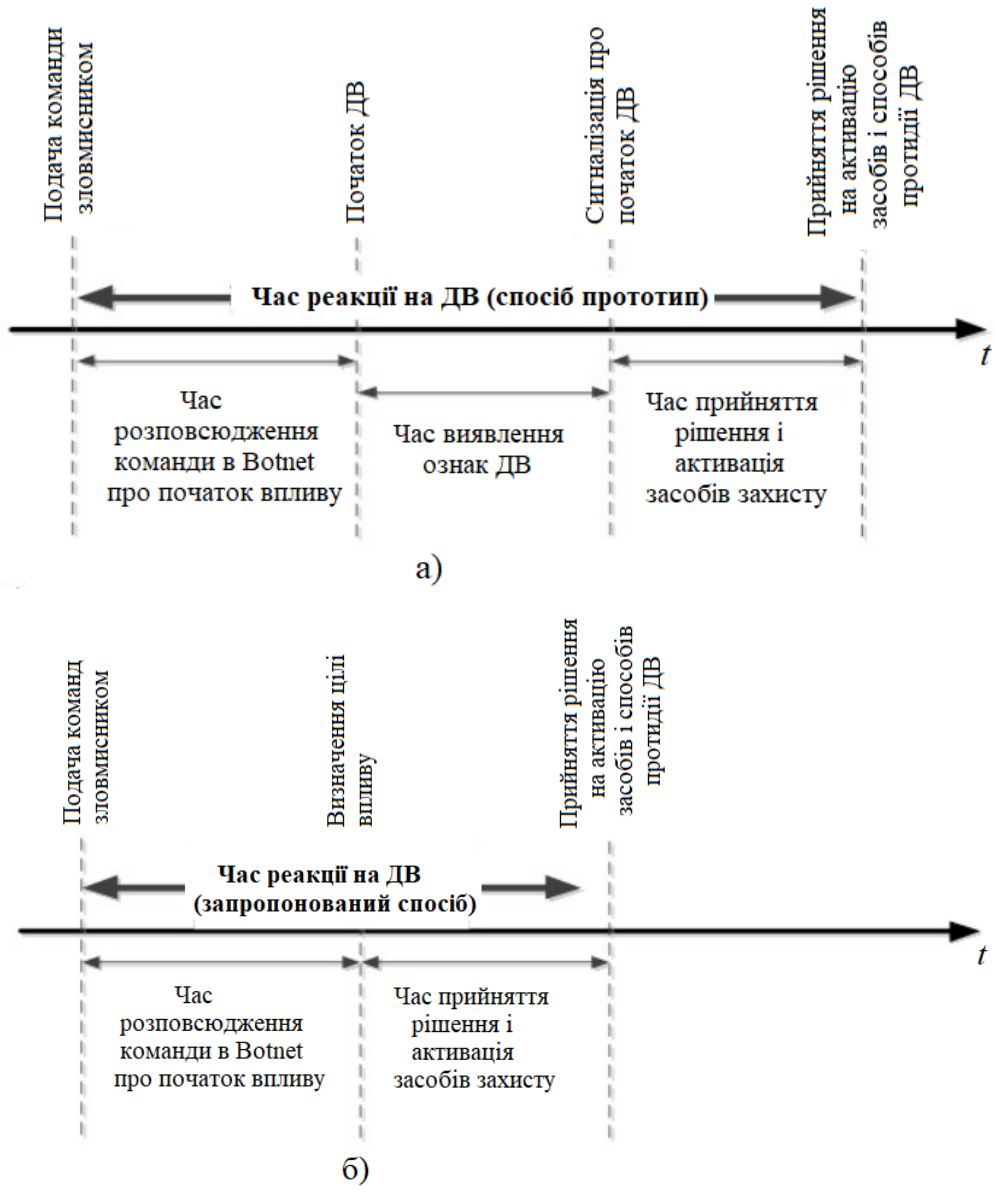


Рисунок 4.4 — Вирішення проблеми раннього виявлення деструктивного впливу Botnet (а) – спосіб прототип; б) -запропонований спосіб)

Запропонований метод класифікації псевдовипадкових послідовностей враховує дискримінуючу здатність статистичних ознак, може бути впроваджений у існуючі засоби запобігання та виявлення витокам інформації з метою усунення зазначених недоліків. Схема використання запропонованого модуля статистичного аналізу у DLP-системах наведено на рис. 4.5. Зашифрований потік даних може передаватися з робочих станцій співробітників, різних інформаційних систем, мережевих сховищ.

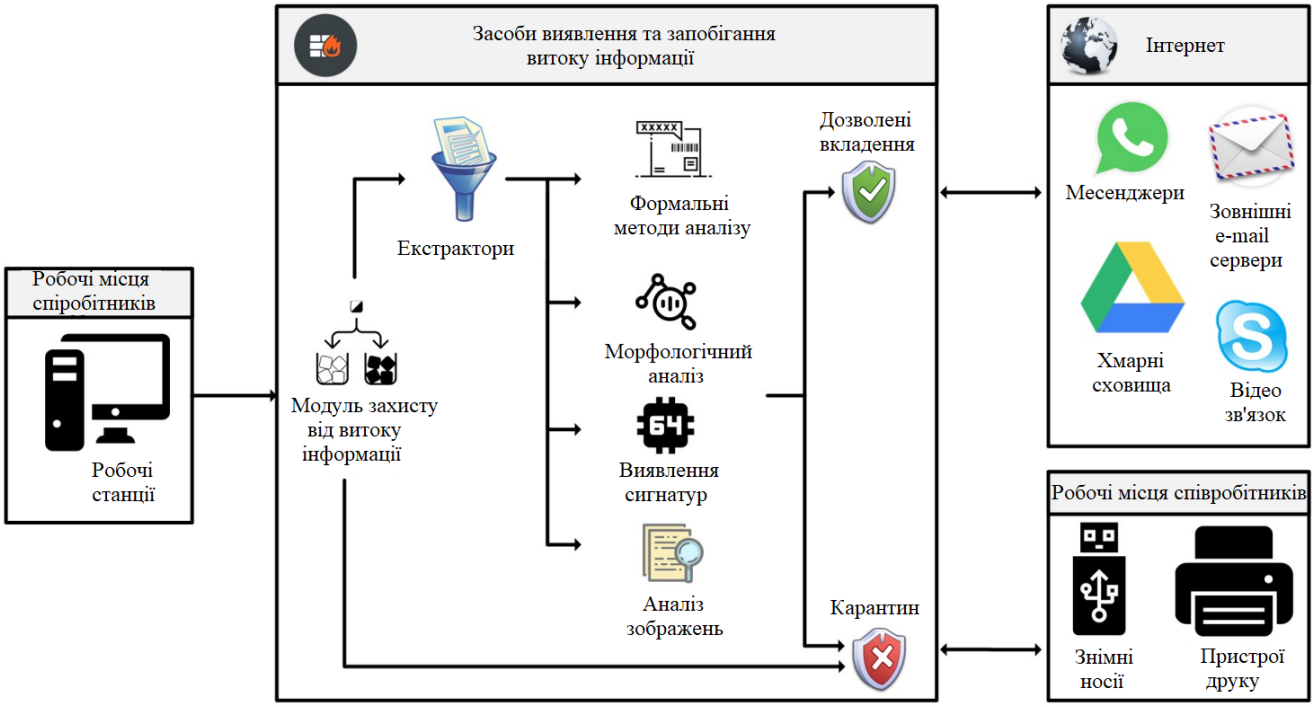


Рисунок 4.5 - Схема впровадження запропонованого модуля статистичного аналізу даних у DLP-системи

При спробі передачі конфіденційної інформації за периметр підприємства по мережевих каналах, при завантаженні даних на знімні носії, модуль статистичного аналізу виконує визначення типу інформації, яка передається. У випадку виявлення зашифрованих послідовностей спрацьовують механізмів захисту, налаштованих відповідно до прийнятої політики безпеки, або здійснюється заборона передачі інформації.

У ході досліджень перенесено запропонований підхід на файли офісних форматів та стислі файли, використано, при цьому, нейронну мережу, алгоритми градієнтного бустингу на основі дерев рішень. Вхідними даними під час проведення класифікації використанні статистичні ознаки: математичне очікування; частота входження підпослідовностей довжиною 4,5,6 байт; розподіл ентропії байт; хешовані значення байтових рядків вкладень середньоквадратичне відхилення значень ентропії байт. Найбільшу точність класифікації послідовностей продемонстрував алгоритм градієтного бустингу на основі дерев

рішень. Запропонований алгоритм класифікації дозволяє здійснювати класифікацію псевдовипадкових послідовностей на основі 100 ознак, що суттєво знижує час проведення класифікації.

На програмному сервері електронної пошти проводиться аналіз вкладень електронної пошти, що знаходиться в межах контрольованого периметру мережі організації. Схема впровадження запропонованого методу класифікації псевдовипадкових послідовностей на сервері пошти наведена на рис. 4.6.

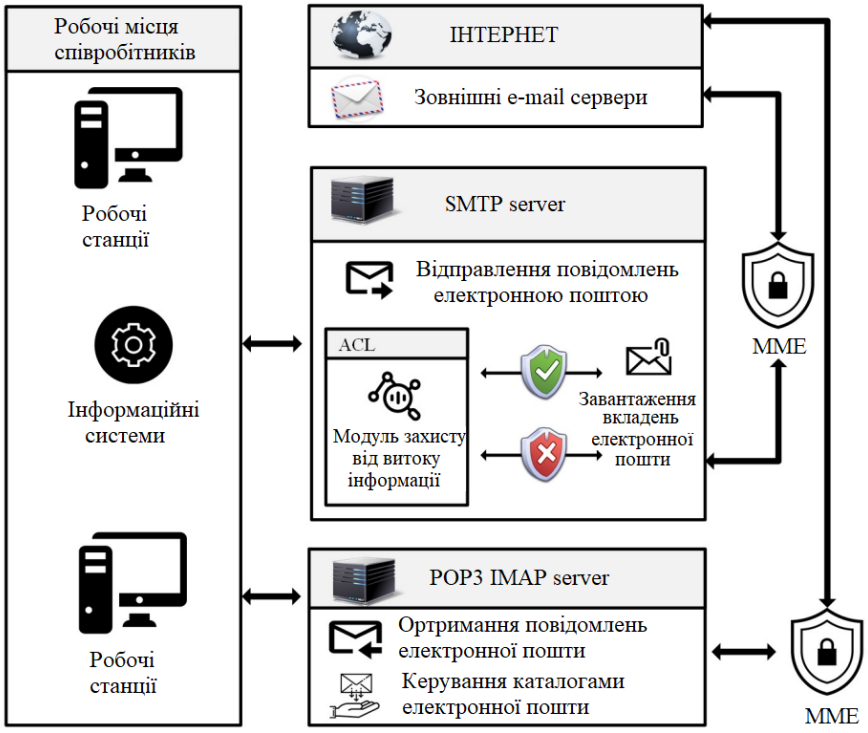


Рисунок 4.6 - Схема впровадження модуля статистичного аналізу даних на сервер електронної пошти методи стиснення даних

Оскільки моделі псевдовипадкових послідовностей і алгоритми класифікації послідовностей застосовуються статистичні ознаки, точність класифікатора залежатиме від розміру аналізованої псевдовипадкової порслідовності.

Для оцінки розробленого алгоритму класифікації та визначення найкращих параметрів класифікатора проведено експерименти над сформованою вибіркою даних. Отримані результати визначення точності класифікації псевдовипадкових

послідовностей від числа використовуваних ознак отриманої моделі, відсортованих за зменшенням дискримінуючої здатності наведені на рис. 4.7.

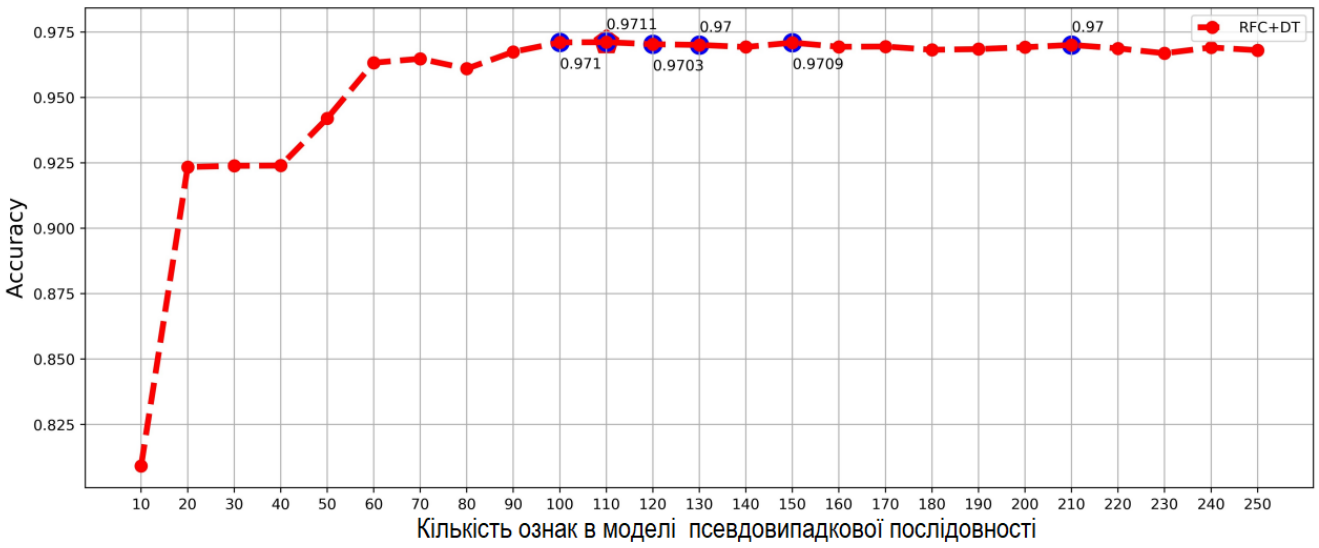


Рисунок 4.7 — Оцінка точності класифікатора від числа ознак моделі псевдовипадкової послідовності

Алгоритм побудови випадкового лісу відноситься до ансамблевих методів, то визначення класу псевдовипадкових послідовностей застосовується процедура голосування яка входить до складу класифікаторів. Таким чином, клас, який набрав більшість голосів, присвоюється аналізованій псевдовипадковій послідовності. Результати визначення найбільш оптимальної глибини дерев та залежність точності класифікації від їх кількості представлені на рис. 4.8 та 4.9 відповідно.

Оскільки алгоритми класифікації псевдовипадкових послідовностей і моделі послідовностей використовуються статистичні ознаки, точність класифікатора залежатиме від розміру псевдовипадкової послідовності.

Результати, від мінімального розміру псевдовипадкової послідовності, оцінки точності класифікатора наведено на рис. 4.10.

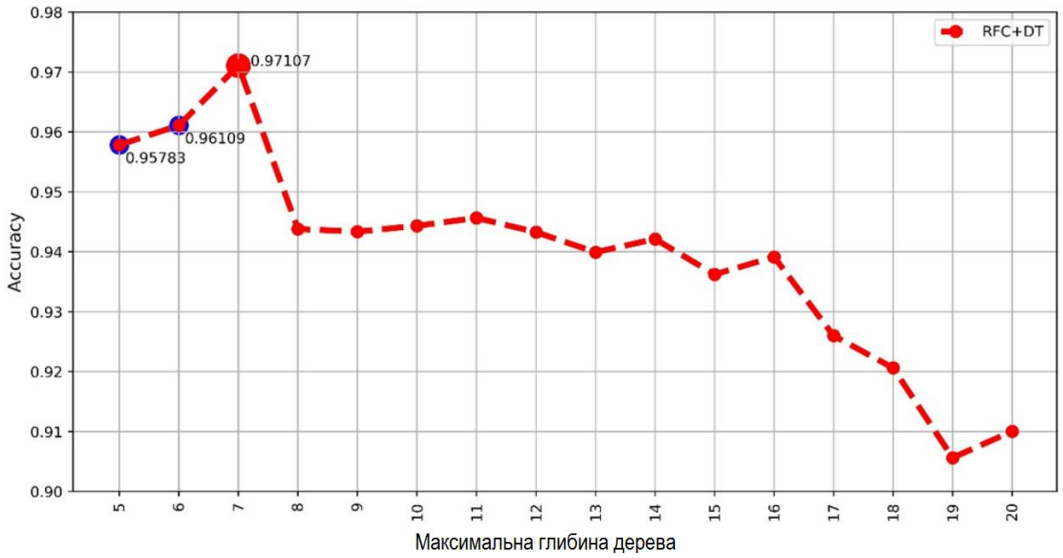


Рисунок 4.8 - Оцінка точності класифікатора від максимальної глибини дерев

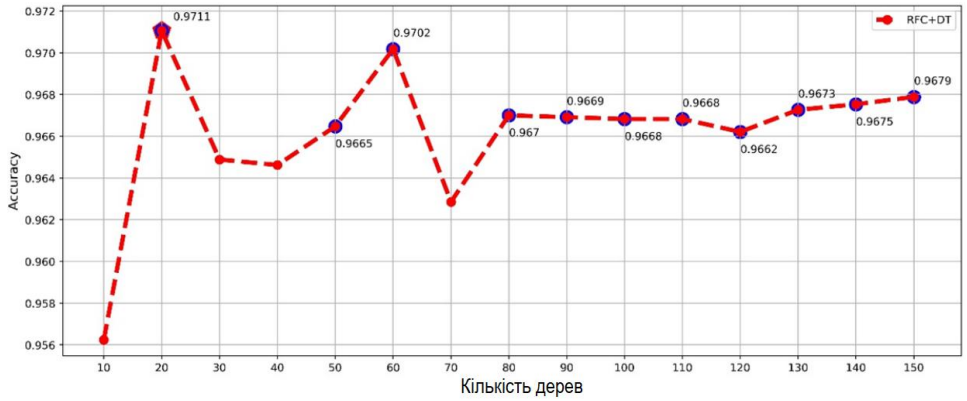


Рисунок 4.9 — Оцінка точності класифікатора від кількості дерев

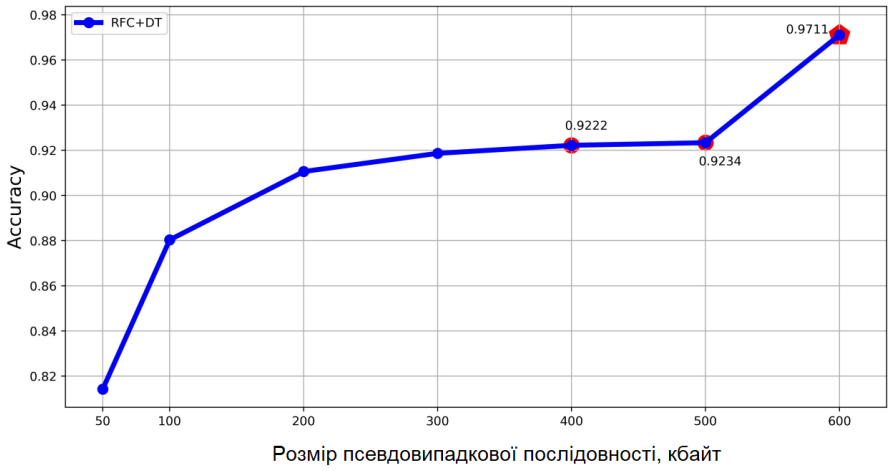


Рисунок 4.10 - Оцінка точності класифікатора від розміру послідовності

Для оцінки ефективності запропонованого методу захисту від витоку конфіденційних даних проведені експерименти з визначення точності бінарної класифікації стислих та зашифрованих даних в залежності від типів вхідних послідовностей, що піддаються процедурам стиснення.

4.3 Висновки

Запропонований метод класифікації псевдовипадкових послідовностей, сформованих алгоритмами стиснення і шифрування інформації, враховує дискримінуючу здатність статистичних ознак даних і його реалізація. Представлено опис, виконано обґрунтування, здійснено пошук основних параметрів класифікатора.

У ході практичної реалізації проведено кількісну оцінку точності класифікації псевдовипадкових послідовностей залежно від параметрів запропонованого класифікатора. Обґрунтовано вибір довжини підпослідовності в дев’ять біт, як значення найбільш раціональнt, що дозволяє досягти класифікації псевдовипадкових послідовностей високої точності та мінімального часу виконання процедури класифікації. Обґрунтовано вибір скануючого оптимального вікна класифікатора розміром в 600 кбайт. Залежно від вимог до точності та швидкості аналізу даних запропоновано два режими роботи: сканування випадково вибраного фрагмента файлу розміром 600 кбайт.

Наведено опис місць впровадження запропонованого алгоритму класифікації псевдовипадкових послідовностей у підсистеми захисту електронної пошти, системи виявлення мережевих атак, засоби запобігання та виявлення витокам інформації. Здійснено порівняльну оцінку запропонованого алгоритму з відомими аналогами в предметній області досліджень. Отримані результати точності класифікації послідовностей дозволяють зробити висновок про досягнення мети магістерського дослідження.

ВИСНОВКИ

Основні результати магістерського дослідження полягають у наступному:

1. На основі проведеного аналізу особливостей функціонування відомих засобів запобігання та виявлення витоку інформації, виявлено обмеження підходів класифікації стиснених та зашифрованих даних, пов'язані з використання заголовків для проведення аналізу файлів та низькою точністю класифікації без використання заголовків.

2. У ході проведених експериментів визначено раціональну довжину підпослідовностей – дев’ять біт, що беруть участь у формуванні простору ознак.

3. Модель псевдовипадкових послідовностей, сформованих алгоритмами стиснення та шифрування даних, дозволила врахувати особливості стиснених та зашифрованих псевдовипадкових послідовностей при поданні в бінарному виді, підпослідовностями довжиною в дев’ять біт.

4. Метод класифікації псевдовипадкових послідовностей, сформованих алгоритмами стиснення та шифрування даних, враховує дискримінуючу здатність статистичних ознак послідовностей, показує більш високу точність класифікації на відміну від відомих аналогів.

5. Проведено оцінку ефективності запропонованих підходів. Отримані значення точності класифікації псевдовипадкових послідовностей перевищують відомі аналоги.

Отримані значення часу та точності класифікації запропонованого алгоритму в порівнянні з існуючими дослідженнями в заданій предметній області дозволяють зробити висновок про досягнення мети магістерського дослідження.

За темою роботи опубліковано 1 теза та 1 наукова стаття.

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ

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

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

3. Богуш, В.М. Криптографічні застосування елементарної теорії чисел / В.М. Богуш, В.А. Мухачов. – К.: ДУІКТ, 2016. – 126 с.

4. Бурячок, В.Л. Інфомаційна та кібербезпека / В.Л. Бурячок, В.Б. Толубко, В.О. Хорошко – К.: ДУТ, 2015р. – 288 с.

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

6. Виростков, Д. Огляд способів і протоколів аутентифікації в веб- додатках [Електронний ресурс] / Д. Виростков. 2017. Режим доступу: https://habr.com/ru/company/dataart/blog/262817

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

8. Голубєв, О.В. Розслідування комп'ютерних злочинів / О.В. Гошубєв – «Запоріз. ін-т муніцип. упр. і держ.», 2016. – 297 с.

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

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

11. Дешко, Л. М., Бондарєва К. Д. Кібербезпека в Україні: національна стратегія та міжнародне співробітництво / Л.М. Дешко, К.Д. Бондарєва - Електронне наукове фахове видання «Аналітичне право». 2018р. 379–382с.

12. 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. – Київ : ВІКНУ, 2022. – Вип. № 74. – С. 73-82.

13. Доктрини «Інформаційної безпеки України» від 25 лютого 2017 року № 47/2017. [Електронний ресурс]. – Режим доступу: https://zakon3.rada.gov.ua/laws/show/47/2017?lang=ru

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

15. Закон України Про захист інформації в інформаційно-телекомунікаційних системах № 80/94-ВР від 05.07.1994 р., [Електронний ресурс]. – Режим доступу: https://zakon.rada.gov.ua/laws/show/

16. Закон України «Про інформацію» - [Електронний ресурс]. – Режим доступу: https://zakon.rada.gov.ua/laws/show/2657-12

17. Закон України Про криптографічний та технічний захист інформації [Електронний ресурс]. – Режим доступу: https://ips.ligazakon.net/document/NT1819

18. Закон України «Про основні засади забезпечення кібербезпеки України» веб-сайт. URL: https://zakon.rada.gov.ua/laws/show/2163-19

19. Захист конфіденційної інформації - персональних даних [Електронний ресурс]. – Режим доступу https://cutt.ly/iuggGRH

20. Клінцв, Л.М. Безпека програм і даних / Л.М. Клінцв – Чернігов: ВСП Чернігівський інститут інформації, бізнесу і права, 2017р. – 81 с.

21. Кормич, Б.А. Інформаційна безпека: організаційно-правові основи: Навч. посібник. / Б.А. Кормич - К.: Кондор, 2015р. - 384 с.

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

23. Кучерявий, Є.І. Методи класифікації зашифрованих даних засобами запобігання та виявлення витоку інформації//Є.І. Кучерявий, В.М. Джулій -

Військова освіта і наука: сьогодення та майбутнє: зб. тез доповідей XIX Міжнародної науково-практичної конференції, м. Київ, 10 листопада 2023 р. Київ: Військовий інститут Київського національного університету імені Тараса Шевченка, 2023. – С. 34.

24. Кучерявий, Є.І. Дослідження методів класифікації зашифрованих та стислих даних засобами виявлення та запобігання витоку конфіленційної інформації / Є.Кучерявий, В.Джулій, І. Муляр. // Вісник Хмельницького національного університету. Технічні науки. – 2023. – № . – С.

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

26. 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

27. 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.

28. 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 Т. Шевченка. – К.: ВІКНУ, 2022. – Вип. №77. – С. 103-117.

29. 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.

30. Логінова, Н. І. правовий захист інформації : навчальний посібник / Н. І. Логінова, Р. Р. Дробожур. – Одеса : Фенікс, 2015. – 264 с.

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

32. Нашинець-Наумова, А.Ю. Інформаційна безпека: питання правового регулювання. – К.: ВД "Гельветика", 2017р. – 168 с.

33. Остапов, С. Е. технологія захисту інформації : навчальний посібник / С. Е. Остапов, С. П. Євсеєв, О. Г. Король. – Х. : Вид. ХНЕУ, 2018р. – 476 с.

34. Петрик, В. Сутність інформаційної безпеки держави, суспільства та особи / В. Петрик. [Електронний ресурс] – Режим доступу: http://www.justinian.com.ua/article.php?id=3222

35. Біленчук, П.Д. Правові засади інформаційної безпеки України: монографія / П.Д. Біленчук, Л.В. Борисова – Харків: 2018р. – 289с.

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

37. Ушатов, В. Проблеми оперативного виявлення і реагування на інциденти інформаційної безпеки / В. Ушатов, О. Сєвєрінов // GLOBAL CYBER SECURITY FORUM. Матеріали першого міжнародного науково-практичного форуму – Х.: ХНУРЕ, 2019р. –104-105с.

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

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

40. An empirical approach towards characterization of encrypted and unencrypted VoIP traffic / P. Choudhury // Multimedia Tools and Applications. – 2020. - т. 79, № 1. - с. 603-631.

41. Casino, F. Hedge: Efficient traffic classification of encrypted and compressed packets / F. Casino, K.-K. R. Choo, C. Patsakis // IEEE Transactions on Information Forensics and Security. - 2019. - т. 14, № 11. - с. 2916-2926

42. Conference on Network and System Security. - Springer. 2020. - с. 42-62.

43. Classification of pseudo-random sequences based on the random forest algorithm / A. A. Spirin [et al.] // Ivannikov Memorial Workshop Proceedings. — 2020. - P. 55-58.

44. Cyber security of critical infrastructures / L. A. Maglaras [и др.] // Ict Express. - 2018. т. 4, № 1. - с. 42-45.

45. EnCoD: Distinguishing Compressed and Encrypted File Fragments / F. De Gaspari// International

46. Europol. (2018). Internet Organised Crime Threat Assessment 2018. [Електронний ресурс]. – Режим доступу : https://www.europol.europa.eu/activities-services/main-reports/internetorganisedcrime-threat-assessment-iocta-2018.

47. Europol. (2018). Public Awareness and Prevention Guides. [Електронний ресурс]. – Режим доступу : https://www.europol.europa.eu/activitiesservices/public-awareness-and-preventionguides

48. Fisher, Tim. (2018) Free and Public DNS Servers. Lifewire. [Електронний ресурс]. – Режим доступу : https://www.lifewire.com/free-and-public-dnsservers

49. Global Cybersecurity Index (GCI) 2018 [Електронний ресурс]. – Режим доступу : https://www.itu.int/en/ITU-D/ Cybersecurity/Documents/draft-18-00706_Global-Cybersecurity-Index-EV5_print_2.pdf.

50. Graham Bartlett, Amjad Inamdar. IKEv2 IPsec Virtual Private Networks: Understanding and Deploying IKEv2, IPsec VPNs, and FlexVPN in Cisco IOS. – Cisco Press, 2016 – 608 с.

51. InfoWatch. Витоки конфіденційних даних / InfoWatch. - 2020. - URL: https:// www . infowatch / analytics / reports / 30708

52. Kozyura V. D., Khoroshko V. O., Shelest M. YE., Tkach YU. M., Usov YA.YU. Kompleksni systemy zakhystu informatsiyi v informatsiyno-

telekomunikatsiynykh systemakh [Complex systems of information protection in information and telecommunication systems]. Nizhyn: FOP Luk'yanenko V.V. TPK «Orkhideya», 2019. 144 p.

53. Le, D. C. Analyzing data granularity levels for insider threat detection using machine learning / D. C. Le, N. Zincir-Heywood, M. I. Heywood // IEEE Transactions on Network and Service Management. - 2020. - т. 17, № 1. - с. 30-44.

54. Mohd, N. Mitigating Insider Threats: A Case Study of Data Leak Prevention / N. Mohd, Z. Yunos // European Conference on Cyber Warfare and Security. — Academic Conferences International Limited. 2020. - с. 599-605.

55. Optimizing feature selection for efficient encrypted traffic classification: A systematic approach / M. Shen // IEEE Network. - 2020. - т. 34, № 4. - с. 20-27.

56. Pokoradi L. Fuzzy logic-based risk assessment. [Електронний ресурс]. – Режим доступу : URL: http:// www.zmka.hu/docs/Volume1/Issue1/pdf/04poko.pdf

57. Self-adaptive system for the corporate area network resilience in the presence of botnet cyberattacks / S. Lysenko // International Conference on Computer Networks. - Springer. 2018. - с. 385-401

58. Shang K., Hossen Z.Applying Fuzzy Logic to Risk Assessment and DecisionMaking // Casualty Actuarial Society, Canadian Institute of Actuaries, Society of Actuaries. November 2017.

ДОДАТОК А
(обовязковий)
Алгоритм отримання ознак на основі моделі
псевдовипадкових послідовностей

Data: ПВП p. класифікатор <K>, <V>
Result: Клас у ПВП p

$F_{Q,V} \leftarrow <>$;
$State \leftarrow <>$;
$M_p \leftarrow \mathbf{Len}(p)$;
for $v \in V$ **do**
    $N_v \leftarrow \mathbf{Len}(v)$;
    $n_v \leftarrow \mathbf{Count}(p,v)$;
    $f_{p,v} = \frac{n_v}{M_p - N_v + 1}$;
    $F_{Q,V} = F_{Q,V} \cup f_{p,v}$;
for $b \in B$ **do**
    $n_b \leftarrow \mathbf{Count}(b,s)$;
    $bytes_p \leftarrow < b,n_b >$;
$F_{Q,E} \leftarrow F_{Q,E} \cup bytes_p$;
$std_p = \mathbf{Std}(bytes_p)$;
$min_p = \mathbf{Min}(bytes_p)$;
$max_p = \mathbf{Max}(bytes_p)$;
$delta_p = max_p - min_p$;
$F_{Q,E} \leftarrow F_{Q,E} \cup < std_p, min_p, max_p, delta_p >$;
$State \leftarrow \mathbf{Next}\,(k)$;
while $State[7] \neq True$ **do**
    if $f_{p,State[2]} \geqslant State[3]$ **then**
        $State \leftarrow \mathbf{NextRight}\,(State)$
    else
        $State \leftarrow \mathbf{NextLeft}\,(State)$
$y_p \leftarrow State[4]$;
return y_p

ДОДАТОК Б
(обовязковий)

Код (лістинг) програмних компонентів системи виявлення витоку в мережах конфіденційної інформації

```
#Алгоритму оцінки джерел активності
#Крок 1 Обчислення суми повідомлень джерела
for (int) i in range (length(list(Sources_Potential_Calculation[URLmessage]))):
 Sources_Potential_Calculation.local[int i,potentialIndex]==0
 if Sources_Potential_Calculation[Type_message][i]= post: else
Sources_Potential_Calculation.local[i,potential_Index]=Sources_Potential_Calculation
[potentialIndex][i]+1
 if SourcesPotentialCalculation[Type_message][i] = comment: else
SourcesPotentialCalculation.loc[i,potentialIndex]=SourcesPotentialCalculation[potenti
alIndex][i]+0.5
 if SourcesPotentialCalculation [messageType][i]= reply to comment: else
SourcesCalculationPotential.loc[(int)i,potentialIndex]=SourcesCalculationPotential [
Indexpotential][i]+0.25
## Крок 2 Розрахунок джерела повідомлень потенціалу
## Розрахунок середнього
count = 0
firstAverage = 0
for int i in range (length(list(SourcesCalculationPotential [URLmessage]))):
 firstAverage = firstAverage + SourcesPotentialCalculation['potentialIndex'][i] else
 count = count + 1
firstAverage = firstAverage / count
# Розрахунок другого середнього
secondAverage = 0
count = 0
for i in range (len(list(SourcesPotentialCalculation['messageURL']))):
 if SourcesPotentialCalculation['potentialIndex'][i]>= firstAverage:
 secondAverage = secondAverage + SourcesPotentialCalculation['potentialIndex'][i]
 count = count + 1
secondAverage = secondAverage / count
```

```
# Формування результату
for i in range (len(list(SourcesPotentialCalculation['messageURL']))):
 if SourcesPotentialCalculation['potentialIndex'][i]<firstAverage:
 SourcesPotentialCalculation.loc[i,'potentialIndex']=0
 elif SourcesPotentialCalculation['potentialIndex'][i]>=firstAverage and
SourcesPotentialCalculation['potentialIndex'][i]<secondAverage:
 SourcesPotentialCalculation.loc[i,'potentialIndex']=1
 elif SourcesPotentialCalculation['potentialIndex'][i]>=secondAverage:
 SourcesPotentialCalculation.loc[i,'potentialIndex']=2
#Алгоритм оцінки джерел активності
Крок 1. Обчислення індексу активності
 if SourcesActivityCalculation['sourseId'][j]==subscriberCount['sourseId'][i]:
SourcesActivityCalculation.loc[j,'subscriberCount']=subscriberCount['subscriberCount']
[i]
urlCounter=0
for i in range (len(list(SourcesActivityCalculation['messageURL']))):
 for j in range (len(list(SourcesActivityCalculation['messageURL']))):
 if i!=j:
 if SourcesActivityCalculation['sourseId'][i]== SourcesActivityCalculation['sourseId'][j]:
 urlCounter=urlCounter+1
urlCounter=list(SourcesActivityCalculation['sourseId'])
urlCounter = len (set (urlCounter))
activityIndex = pd.DataFrame({
 'sourseId':[],
 'activityIndex': [], })
for i in range(len(list(SourcesActivityCalculation['sourseId']))):
 activityIndex.loc[i,'sourseId']= SourcesActivityCalculation['sourseId'][i]
activityIndex.loc[i,'activityIndex']=SourcesActivityCalculation['likesCount'][i]+Sources
ActivityCalcul
ation['commentCount'][i]+SourcesActivityCalculation['repostCount'][i]
for i in range(len(list(activityIndex['sourseId']))):
 if SourcesActivityCalculation['subscriberCount'][i]!=0:
activityIndex.loc[i,'activityIndex']=activityIndex['activityIndex'][i]/SourcesActivityCalc
ulation['subscr
```

```
iberCount'][i]
 else:
 activityIndex.loc[i,'activityIndex']=activityIndex['activityIndex'][i]/1
 activityIndex.loc[i,'activityIndex']=activityIndex['activityIndex'][i]/urlCounter
#Крок 2 Обчислення індексу перегляду джерела
viewIndex = pd.DataFrame({
 'sourseId':[],
 'viewIndex': [], })
for i in range(len(list(SourcesActivityCalculation['sourseId']))):
 viewIndex.loc[i,'sourseId']= SourcesActivityCalculation['sourseId'][i]
 viewIndex.loc[i,'viewIndex']= SourcesActivityCalculation['viewCount'][i]
for i in range(len(list(viewIndex['sourseId']))):
 if SourcesActivityCalculation['subscriberCount'][i]!=0:
viewIndex.loc[i,'viewIndex']=viewIndex['viewIndex'][i]/SourcesActivityCalculation['su
bscriberCount'
][i]
 else:
 viewIndex.loc[i,'viewIndex']=viewIndex['viewIndex'][i]/1
 viewIndex.loc[i,'viewIndex']=viewIndex['viewIndex'][i]/urlCounter
 # Крок 3 Обчислення індексу впливу джерела
 impactIndex = pd.DataFrame({
 'sourseId':[],
 'impactIndex': [], })
for i in range(len(list(SourcesActivityCalculation['sourseId']))):
 impactIndex.loc[i,'sourseId']= SourcesActivityCalculation['messageURL'][i]
 impactIndex.loc[i,'impactIndex']=
activityIndex['activityIndex'][i]+viewIndex['viewIndex'][i]
#Формування результату
InfluenceObjectSorting = pd.DataFrame({...})
```

ДОДАТОК В
(обовязковий)
Копії наукових публікацій

ВІЙСЬКОВИЙ ІНСТИТУТ
КИЇВСЬКОГО НАЦІОНАЛЬНОГО УНІВЕРСИТЕТУ
ІМЕНІ ТАРАСА ШЕВЧЕНКА

ЗБІРНИК
ТЕЗ ДОПОВІДЕЙ

XIX Міжнародної науково-практичної конференції

«Військова освіта і наука: сьогодення та майбутнє»

10 листопада 2023 року

Київ – 2023

ЗМІСТ

к.т.н., доц. Джулій В.М. (ХмНУ)
Кучерявий Є.І. (ХмНУ)

МЕТОДИ КЛАСИФІКАЦІЇ ЗАШИФРОВАНИХ ДАНИХ ЗАСОБАМИ ЗАПОБІГАННЯ ТА ВИЯВЛЕННЯ ВИТОКУ ІНФОРМАЦІЇ

Проведений аналіз досліджень предметної області та об'єкта дослідження дозволяє висунути припущення про наявність у стислих та зашифрованих даних статистичних особливостей. У разі справедливості висунутої гіпотези в результаті проведених досліджень можливо реалізувати модель псевдовипадкових послідовностей, сформованих алгоритмами стиснення та шифрування інформації і запропонувати метод захисту від витоків переданої інформації на основі поділу типів даних.

Розглянуті методи класифікації переданих стислих і зашифрованих даних дозволяють висунути вимоги, яким повинні задовольняти засоби захисту запобігання та виявлення витоку конфіденційної інформації: безперервність роботи системи в часі; оперативність аналізу переданих даних; незалежність від типу контейнера даних.

До методу безпеки даних від витоку даних, пред'являються наступні вимоги: використання статистичних методів, незалежних від характеристик контейнерів передачі та зберігання даних; класифікація лише підозрілих послідовностей; формат аналізованих даних не важливий, дані надходять в бінарному вигляді; точність класифікації стислих та зашифрованих даних має досягати максимального значення; можливість протидії загрозам безпеки корпоративних мереж підприємства, також протидія та виявлення botnet мережам; час виконання класифікації має досягати до мінімуму.

Здійснено формальну постановку задачі, визначено мету, проведено аналіз об'єкта та предмета дослідження. Відсоток інцидентів порушення безпеки, конфіденційних даних, пов'язаних із витоком інформації, причиною яких є внутрішні зловмисники, склав понад 78%, що підтверджує актуальність дослідження. Проведено аналіз вразливостей та загроз DLP–систем та засобів захисту даних, визначено недоліки та переваги використовуваних підходів класифікації стислих та зашифрованих даних.

Обґрунтовано вибір статистичних методів проведення аналізу переданих даних для побудови класифікатора, сформульована наукова задача магістерського дослідження у формальному виді. Показано практичну проблему, яка полягає в низькій точності класифікації стислих і зашифрованих псевдовипадкових послідовностей.

УДК 004.056:621.397.3:004.942

ЄВГЕН КУЧЕРЯВИЙ
ORCID https://orcid.org/0009-0004-1867-6241
Хмельницький національний університет
e-mail: gorix2019@gmail.com
ВОЛОДИМИР ДЖУЛІЙ
Хмельницький національний університет
ORCID http://orcid.org/0000-0003-1878-4301
e-mail: dzhuliivm@khmnu.edu.ua
ІГОР МУЛЯР
Хмельницький національний університет
ORCID http://orcid.org/0000-0002-6659-605X
e-mail: muliariv@khmnu.edu.uaK

ДОСЛІДЖЕННЯ МЕТОДІВ КЛАСИФІКАЦІЇ ЗАШИФРОВАНИХ ТА СТИСЛИХ ДАНИХ ЗАСОБАМИ ВИЯВЛЕННЯ ТА ЗАПОБІГАННЯ ВИТОКУ КОНФІЛЕНЦІЙНОЇ ІНФОРМАЦІЇ

Розглянуті методи класифікації переданих зашифрованих та стислих даних, дозволяють висунути вимоги, яким повинні задовольняти засоби захисту запобігання та виявлення витоку конфіденційної інформації: безперервність роботи системи в реальному часі; оперативність аналізу переданих даних; незалежність від типу контейнера даних.

Здійснено формальну постановку задачі, визначено мету, проведено аналіз об'єкта та предмета дослідження. Відсоток інцидентів порушення безпеки, конфіденційних даних, пов'язаних із витоком інформації, причиною яких є внутрішні зловмисникм, склав понад 78%, що підтверджує актуальність дослідження. Проведено аналіз вразливостей та загроз DLP–систем та засобів захисту даних, визначено недоліки та переваги використовуваних підходів класифікації зашифрованих та стислих даних.

Обґрунтовано вибір статистичних методів проведення аналізу переданих даних для побудови класифікатора, сформульована задача дослідження у формальному виді. Показано практичну проблему, яка полягає в низькій точності класифікації зашифрованих та стислих псевдовипадкових послідовностей.

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

YEVHEN KUCHERYAVYI, VOLODYMYR DZHULIY, IHOR MULIAR
Khmelnytsky National University

RESEARCH OF METHODS OF CLASSIFICATION OF ENCRYPTED AND COMPRESSED DATA BY MEANS OF DETECTING AND PREVENTING LEAKAGE OF CONFIDENTIAL INFORMATION

Abstract. The considered methods of classification of transmitted encrypted and compressed data make it possible to set requirements that must be met by means of protection for the prevention and detection of leakage of confidential information: continuity of system operation over time; speed of analysis of transferred data; independence from the type of data container.

The following requirements apply to the method of data security against data leakage: the use of statistical methods independent of the characteristics of data transfer and storage containers; classification of only suspicious sequences; the format of the analyzed data is not important, the data is received in binary form; the accuracy of the classification of compressed and encrypted data should reach the maximum value; the possibility of countering threats to the security of the company's corporate networks, as well as countering and detecting botnet networks; the classification execution time should reach a minimum.

The formal formulation of the problem was carried out, the goal was determined, and the object and subject of the

research were analyzed. The percentage of incidents of security breaches, confidential data related to information leakage, which are caused by internal attackers, amounted to more than 78%, which confirms the relevance of the study. An analysis of the vulnerabilities and threats of DLP systems and data protection tools was carried out, the disadvantages and advantages of the used approaches for classifying encrypted and compressed data were determined.

The choice of statistical methods for the analysis of the transmitted data for the construction of the classifier is justified, the scientific task of the research is formulated in a formal form. A practical problem is shown, which consists in the low accuracy of classification of compressed and encrypted pseudo-random sequences equal to 0.95 and the use of headers of the specified files.

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

Вступ

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 (внутрішньому порушнику) слабо розвинені, в більшості документів, що регламентують політику безпеки кофіденційним даним компанії, містяться постулати про відсутність інсайдера, що тягне, в даному випадку, зростання ймовірності порушення інформаційної безпеки даних, що захищаються [1,3].

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 [4,13].

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 [2,5,14].

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 [3,8,9].

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 оцінци 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 [7,11,13].

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 [5,6,10].

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 [10,13].

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 [10,13,14].

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.

Постановка задачі

Інформаційні технології, на теперішній час, 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.

Незважаючи на удосконалення механізмів захисту від кіберзагроз, розвиток засобів захисту конфіденційної даних, 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, але здійснювати передачу конфіденційної інформації за контрольований інформаційний периметр компанії.

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 [1,6,7].

Відсутність у корпоративній моделях 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.

Забезпечення інформаційної безпеки конфіденційних даних та призупинення дій внутрішнього порушника здійснюється за допомогою, в основному, організаційних заходів. Виконується тестування, перевірка фактів їхньої біографії, відбір кандидатів, протягом проміжку часу можуть змінитися багато факторів, один з яких - лояльність співробітника [1,2,11].

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.

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.

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 [10,11,13]. Атаки на такі критичні інформаційні системи включають проникнення в мережу організації та встановлення шкідливого програмного забезпечення, які можуть розкрити конфіденційну інформацію, змінити поведінку конкретного технічного обладнання. 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 [7,10,11].

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.

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

Основна частина

Корпоративні системи зберігання та обробки даних призначені для виконання процедур зберігання, передачі, перетворення інформації різного ступеня секретності, типу, а також персональні дані співробітників, користувачів компанії, а у випадку державної установи також дані великої кількості громадян. Даний факт дозволяє розглядати корпоративні системи компанії, як інформаційні системи зберігання, обробки персональних даних. У зазначених інформаційних системах під зловмисником розуміється фізична особа, яка навмисно чи випадково вчиняє дії, які є наслідком порушення безпеки персональних даних співробітників при обробці інформації технічними засобами в інформаційно-пошукових системах персональних даних [6,7,13].

Залежно від наявності у зловмисників легітимного доступу до інформаційно-пошукової системи, системи поділяються: зловмисники, які мають доступ до корпоративної мережі організації передачі даних, включаючи, також, користувачів інформаційно-пошукової системи, що реалізують загрози безпосередньо у системі – внутрішні зловмисники; зловмисники, які не мають доступу до корпоративної мережі компанії передачі даних, реалізують атаки, загрози із мереж міжнародного інформаційного обміну або зовнішніх мереж зв'язку загального користування – зовнішні порушники.

Можливості внутрішнього зловмисника на теперішній час дуже великі і становлять серйозну безпеку для конфіденційної інформації, враховуючи, при цьому, широкий спектр програмно-апаратний засобів, що реалізують процедури стиснення чи шифрування інформації.

DLP–системи дозволяють проводити аналіз переданої інформації для всіх груп у використання клієнт-серверної архітектури інформаційно-пошукової системи в корпоративній мережі організації, у випадку якщо порушник не використовує методи стиснення та шифрування переданих даних. В даному випадку внутрішній зловмисник здатний реалізувати витік конфіденційної інформації в обхід існуючих політик безпеки та систем захисту, що унеможливлює реєстрацію події безпеки інформації, і також, розслідування інциденту безпеки даних.

Вразливість інформаційно-пошукової системи в корпоративній мережі організації передачі даних є слабким місцем (недоліком) у прикладному, системному програмному забезпеченні автоматизованої інформаційно-пошукової системи, яка використовується для реалізації атак, загроз безпеці конфіденційних даних. Причинами виникнення вразливості є [10,13]: навмисні дії внесення вразливості в ході розробки та проектування програмно-апаратного забезпечення; помилки при розробці та проектуванні програмно апаратного забезпечення; несанкціоноване використання та впровадження неврахованих програм з подальшим витрачанням ресурсів (захоплення оперативної пам'яті, завантаження процесора); неправильні налаштування програмно-апаратного забезпечення, неправомірна зміна режимів роботи програм та пристроїв; несанкціоновані ненавмисні дії співробітників, що призводять до виникнення в системі вразливостей; збої в роботі програмно-апаратного забезпечення (викликані виходом з ладу апаратних елементів, збоями в електроживленні, зовнішніми впливами електромагнітних полів); впровадження шкідливого програмного забезпечення, що створюють вразливості у програмно-апаратному забезпеченні.

Для зазначених атак, загроз інформаційній безпеці необхідно визначити значення відповідних показників рівня небезпеки. Відповідно до [2,4], показники рівня небезпек даних представлені у таблиці 1. Внутрішній зловмисник має локальний доступ до інформації організації, оскільки має легітимний доступ до корпоративної

мережі підприємства, також, можливе впровадження шкідливого програмного забезпечення, яке приховано функціонує в корпоративній мережі організації (botnet агенти).

Таблиця 1

Значення рівня показників небезпеки інформації

Показник рівня небезпеки інформації	Найменування показника небезпеки	Значення показників рівня небезпеки
Тип доступу (td)	Віддалений	3
	Фізичний	1
	Локальний	2
Рівень складності (dl)	Помірний	4
	Середній	2
	Підвищений	3
	Високий	1
Значимість інформаційних компонентів, ресурсів (rc)	Низька	1
	Середня	2
	Висока	3

Рівень складності реалізації небезпеки даних є помірним (найнижчим) з представлених, що обумовлюється простою схемою реалізації загрози передачі даних в стислому чи зашифрованому виді як внутрішнім зловмисником так і шкідливим програмним забезпеченням. Значимість інформаційних ресурсів обрана середньою. Показник рівня небезпеки загрози конфіденційним даним складається із суми зпоказників і становить 8, відповідно до виразу (1):

$$(W) = td(2) + dl(4) + rc(2) = 8 \quad (1)$$

Відповідно до методики [2,7], необхідно визначити рівень небезпеки загроз безпеки даних за таблицею 2.

Значення отримане у виразі (1) відповідає високому рівню загрози інформаційній безпеці даних, що підтверджує, в даній ситуації, актуальність теми досліджень. На рис. 1 представлена схема процесу витоку даних в стислому або зашифрованому вигляді за контрольований периметр компанії, реалізована внутрішнім порушником.

Таблиця 2

Рівень небезпеки загроз безпеки даних

Рівень загроз безпеки даних	Діапазон значень
Низький	$W = 4$
Середній	$5 \le W \le 7$
Високий	$W = 8$

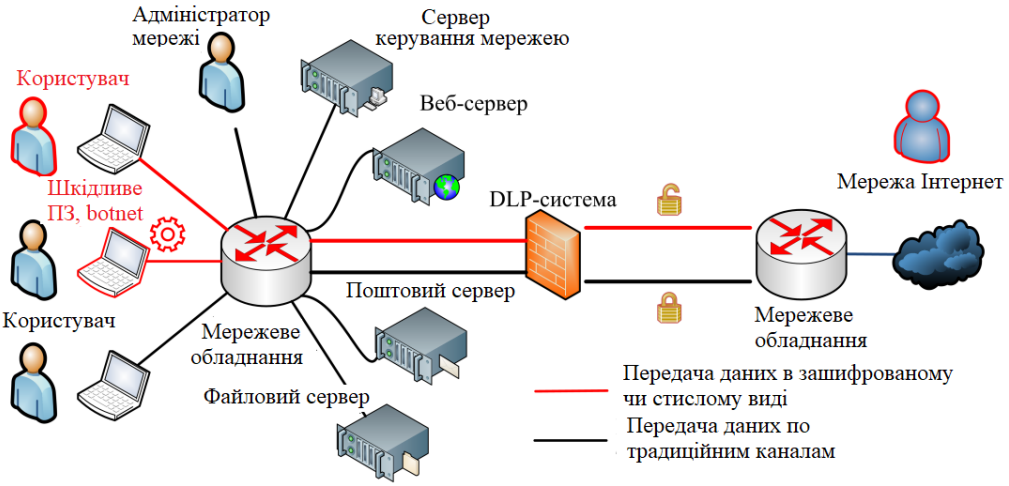


Рис. 1. Схема процесу витоку даних, реалізована внутрішнім порушником

Передача даних може здійснюватися стандартними способами та засобами, так і за допомогою засобів стиснення та шифрування інформації. У разі використання засобів стиснення та 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.

З метою обґрунтування актуальності завдання класифікації зашифрованих, стислих та відкритих даних було проведено аналіз досліджень в області захисту інформації. За даними експертно-аналітичного центру SafeNet в 2021р. в Україні близько 78% зафіксованих витоків конфіденційних даних сталися з вини внутрішніх зловмисників, близько 76% з них були навмисними [4,5]. Статистика зафіксованих витоків інформації за 2021р. представлена на рис2.

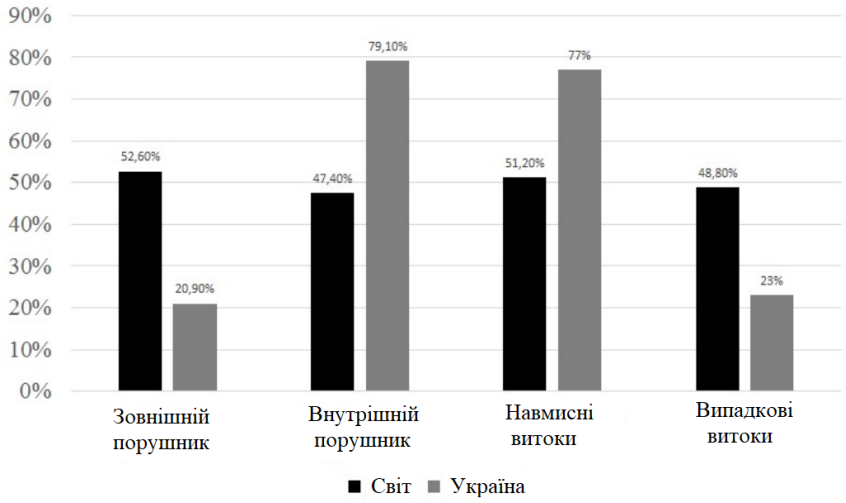


Рис. 2. Статистика зафіксованих витоків інформації за 2021р.

Загрози від внутрішніх зловмисників є важкоусувними та найбільш небезпечними, у тому числі при використанні засобів стиснення та шифрування інформації, що дозволяє, при цьому, не порушуючи політик безпеки підприємства відправляти конфіденційні дані за периметр організації. Ексфільтрація конфіденційної інформації за периметр компанії (крадіжка даних) може здійснюватися широким колом користувачів, що належать до внутрішніх зловмисників. Найбільша кількість витоків конфіденційної інформації спостерігається в великих організаціях в корпоративних мережах, що здійснюють обробку персональної інформації, а також з мереж державних установ та організацій. Основними джерелами витоків інформації - системи та сервіси електронної пошти, що мають доступ до мережі Інтернет.

Високий відсоток витоків конфіденційної інформації може бути обумовлений наявністю недоліків та вразливості у існуючих засобах захисту даних, також в DLP-системах. Сучасні засоби захисту даних не виявляють канали витоку інформації, сформованих шляхом використання стиснення та шифрування даних, що робить такий захист малоефективними забезпеченням конфіденційності інформації.

Методи, що розробляються, класифікації стислих і зашифрованих даних, повинні забезпечити підвищення точність класифікації, застосовуватися в архітектурі клієнт-сервер на серверній стороні. Даний підхід дозволить скоротити час, що витрачається на перевірку послідовностей, скоротити обчислювальні ресурси, необхідні для проведення аналізу даних. Підсистема захисту даних від витоку інформації повинна бути інваріантною щодо форматів представлення даних та будь-якої іншої службової інформації. Незалежно від контейнера інформації, що передається за периметр підприємства, повинен визначатися тип даних, стислий або зашифрований. Даний підхід дозволить у режимі реального часу проводити класифікацію потенційно небезпечної інформації та своєчасно реагувати на інциденти безпеки інформації.

Для виявлення передачі стиснутих чи зашифрованих даних, які містять конфіденційну інформацію та визначення джерела несанкціонованого поширення даних, необхідно розробити алгоритм класифікації псевдовипадкових послідовностей. Псевдовипадкові послідовності - файли, що проходять тести NIST на випадковість, розподіл байт в яких підпорядковується рівномірному закону розподілу, їх ентропія має значення більше 7,5. Для вирішення даної задачі необхідно провести порівняльний аналіз засобів, методів, технологій класифікації відкритих, стислих, зашифрованих даних [6,13].

Для запобігання витокам конфіденційних даних застосовують технічні та організаційні заходи, які різняться застосуванням апаратних та програмних засобів. Найбільш поширеним програмним засобом запобігання витоку конфіденційної інформації є DLP-системи, що здійснюють аналіз потоків даних на предмет наявності конфіденційної інформації. Зазначені методи виконують аналіз потоків інформації на предмет наявності фраз, певних слів, регулярних виразів, здійснюють оцінку контексту переданих даних, способів і службових характеристик протоколів передачі інформації. Однак, існують засоби обходу подібних механізмів захисту, наприклад стиснення або шифрування даних [1,7,10].

Класифікація методів, які застосовують DLP-системи для виявлення конфіденційної інформації, представлена на рис. 3. Методи можна розділити на дві групи: контекстні та контентні. Контекстні методи - орієнтовані на конкретні протоколи передачі інформації та технології, враховують різні ознаки та факти, що супроводжують процеси обміну даними, адреси одержувачів, джерел, розмір пакета, номер порту програмного забезпечення, що здійснює передачу. Контентні методи здійснюють пошук зліпків, цифрових відбитків, регулярних виразів, здійснюють аналіз переданої інформації, також службової інформації.

До контентних методів відносяться ентропійні підходи - виконують підрахунок ентропії блоків інформаціїх різної довжини, проте не застосовні для класифікації стислих і зашифрованих даних. Статистичні методи і тести на випадковість становлять інтерес для досліджень, так як до алгоритмів шифрування пред'являються певні стандарти по розсіюванню вихідних даних, до алгоритмів стиснення подібні вимоги не висуваються.

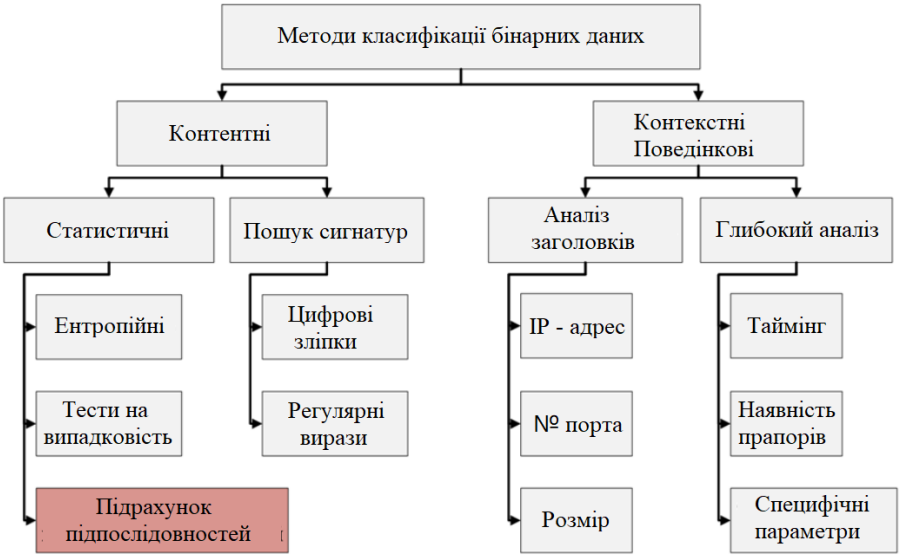


Рис. 3. Класифікація методів, що використовуються в DLP-системах

Оскільки співробітник підприємства має можливість шифрувати конфіденційну інформацію та передати поза периметр організації, задача ідентифікації стислих та шифрованих даних є актуальною. Зі зростанням популярності криптографічних протоколів та їх впровадженні в телекомунікаційні Інтернет мережі деякі засоби безпеки даних, засновані на глибокому аналізі пакетів, перестають достовірно працювати, не можуть виділити ознаки, на яких здійснювалося детектування потенційно небезпечні дії. Відзначається, також збільшену складність проведення аналізу бінарних файлів, з огляду на збільшення їх кількості, поділяють бінарні послідовності на три класи: з низькою ентропією (нестислі медіа-файли), середньою ентропією (структури даних, текст, виконувані файли), високоентропійні – зашифровані та стислі дані. Для класифікації з трьох класів послідовностей пропонують використовувати наступні ознаки: ентропія Шеннона, середнє значення байт у послідовності, вага Хеммінга, ψ - квадрат. Для проведення класифікації використовується алгоритм *k*-найближчих сусідів, точність класифікації відкритих та зашифрованих даних становила близько 96%.

Системи, що здійснюють класифікацію інформації, дозволяють зробити перший крок до виявлення шкідливих дій та вторгнень користувачів в системи. Спочатку системи захисту класифікувалися двома способами: на основі глибокого аналізу властивостей IP- пакета без аналізу даних і на основі аналізу заголовка пакета (номер порту, IP-адреса). Пропонується використовувати на основі дерева рішень методи машинного навчання, генетичних алгоритмів та адаптивного бустингу. Отримані результати свідчать про можливість класифікації відкритих (незашифрованих) та зашифрованих даних з точністю більше 0,96.

Проведений аналіз досліджень предметної області та об'єкта дослідження дозволяє висунути припущення про наявність у стислих та зашифрованих даних статистичних особливостей. У разі справедливості висунутої гіпотези в результаті проведених досліджень можливо реалізувати модель псевдовипадкових послідовностей, сформованих алгоритмами стиснення та шифрування інформації і запропонувати метод захисту від витоків переданої інформації на основі поділу типів даних.

Висновки з даного дослідження і перспективи подальших розвідок у даному напрямі

Розглянуті методи класифікації переданих зашифрованих та стислих даних дозволяють висунути вимоги, яким повинні задовольняти засоби захисту запобігання та виявлення витоку конфіденційної інформації: безперервність роботи системи в часі; оперативність аналізу переданих даних; незалежність від типу контейнера даних.

До методу безпеки даних від витоку даних, пред'являються наступні вимоги: використання статистичних методів, незалежних від характеристик контейнерів передачі та зберігання даних; класифікація лише підозрілих

послідовностей; формат аналізованих даних не важливий, дані надходять в бінарному вигляді; точність класифікації стислих та зашифрованих даних має досягати максимального значення; можливість протидії загрозам безпеці корпоративних мереж підприємства, також протидія та виявлення botnet мережам; час виконання класифікації має досягати до мінімуму.

Таким чином, задача у формальному вигляді може бути визначено виразом (2):

$$\begin{aligned} & F\left(x_i, y_i\right)=1, i=j, \\ & i, j \in Y=(0,1,2), \\ & t \rightarrow \min, \\ & Accuracy \rightarrow \max, \end{aligned} \tag{2}$$

де x_i – файл, що аналізується, y_i - клас файла x_i , $i, j \in Y$ – класи даних: відкриті, зашифровані, стислі.

Здійснено формальну постановку задачі, визначено мету, проведено аналіз об'єкта та предмета дослідження. Відсоток інцидентів порушення безпеки, конфіденційних даних, пов'язаних із витоком інформації, причиною яких є внутрішні зловмисникм, склав понад 78%, що підтверджує актуальність дослідження. Проведено аналіз вразливостей та загроз DLP–систем та засобів захисту даних, визначено недоліки та переваги використовуваних підходів класифікації зашифрованих та стислих даних.

Обґрунтовано вибір статистичних методів проведення аналізу переданих даних для побудови класифікатора, сформульована наукова задача дослідження у формальному виді. Показано практичну проблему, яка полягає в низькій точності класифікації стислих і зашифрованих псевдовипадкових послідовностей дорівнює 0,95 та використання заголовків зазначених файлів.

Літертура

1. 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 КНУ ім. Тараса Шевченка. – К.: ВІКНУ, 2020. – Вип. №68. – С. 53-64.

2. Ленков С.В. Інформаційно-аналітична системи прогнозування вразливостей та загроз інформаційної безпеки / С.В. Ленков, В.М. Джулій, О.В. Мірошніченко, В.О. Браун, С.І. Прохорський. // Збірник наукових праць Військового інституту КНУ ім. Тараса Шевченка.– К.: ВІКНУ, 2023. – Вип. №79. – С. 114-127.

3. Модель потоку текстових повідомлень тематичних інтернет-ресурсів системи прогнозування інформаційної безпеки / В. Джулій, Н. Петляк, Ю. Хмельницький, О. Пахар. // Вісник Хмельницького національного університету. Технічні науки. – 2022. – № 5. – С. 294-300.

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

5. Ленков С.В. Методы и средства защиты информации. В 2-х томах /С.В. Ленков, Д.А. Перегудов, В.А. Хорошко – К: Арий, 2008. – 464с

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

7. 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 КНУ ім. Тараса Шевченка. – К.: ВІКНУ. – 2017. – Вип. № 56. – С.124-132

8. Інформаційно-ознакова модель шкідливої інформації в соціальних мережах / І.В. Муляр, В.М. Джулій, В.М. Пічура, О.О Зацепіна. // Вимірювальна та обчислювальна техніка в технологічних процесах № 3 (2022). – С.73–78.

9. Модель потоку текстових повідомлень тематичних інтернет-ресурсів системи прогнозування інформаційної безпеки / В.М. Джулій, Ю.В. Хмельницький, Н.С. Петляк, О.В. Пахар. // Вісник Хмельницького національного університету. Технічні науки. 2022. № 5. С. 294-300с.

10. Контроль додатків інтернет-трафіка комп’ютерних мереж методами машинного навчання. / Джулій, В.М., Кльоц Ю.П., Муляр І.В., Жилевич М.Л., Джулій А.В. // Вісник Хмельницького національного університету.

Технічні науки. 2021. № 5. С. 22-26.

11. 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 КНУ ім. Тараса Шевченка. – К.: ВІКНУ, 2022. – Вип. №74. – С. 73-82.

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

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

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

References

1. Lenkov S.V. 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 KNU im/ Tarasa Shevchenka. – K.: VIKNU, 2020. – №68. – pp. 53-64.

2. Lenkov S.V. Informatsiino-analitychna systemy prohnozuvannia vrazlyvostei ta zahroz informatsiinoi bezpeky / S.V. Lenkov, V.M. Dzhulii, O.V. Miroshnichenko, V.O. Braun, S.I. Prokhorskyi // Zbirnyk naukovykh prats Viiskovoho instytutu KNU im/ Tarasa Shevchenka. – K.: VIKNU, 2023. –№79. – pp. 114-127.

3. Model potoku tekstovykh povidomlen tematychnykh internet-resursiv systemy prohnozuvannia informatsiinoi bezpeky / V. Dzhulii, N. Petliak, Yu. Khmelnytskyi, O. Pakhar // Herald of Khmelnytskyi National University. Technical sciences. – 2022. – № 5. – pp. 294-300.

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

5. Metodyy sredstva zashchyty ynformatsyy. V 2-kh tomakh / S.V. Lenkov, D.A. Perehudov, V.A. Khoroshko – K: Aryi, 2008. –464s.

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

7. AnalIz Isnuyuchih metodiv ta algoritmiv viyavlennya atak v bezdrotovih merezhah peredachI danih / S.V. Lenkov, V.M. Dzhuliy, N.M. Bernaz, S.O. Bozhuk // Zbirnyk naukovykh prats Viiskovoho instytutu KNU im/ Tarasa Shevchenka. – K.: VIKNU. 2017. – Vip. № 56. – p.124-132

8. Informatsiino-oznakova model shkidlyvoi informatsii v sotsialnykh merezhakh/ I.V. Muliar, V.M. Dzhulii, V. M. Pichura, O.O Zatsepina – Vymiriuvalna ta obchysliuvalna tekhnika v tekhnolohichnykh protsesakh. – № 3 (2022) –S. 73–78.

9. Model potoku tekstovykh povidomlen tematychnykh internet-resursiv systemy prohnozuvannia informatsiinoi bezpeky / V.M. Dzhulii, Yu.V. Khmelnytskyi, N.S. Petliak, O.V. Pakhar // Herald of Khmelnytskyi National University. Technical sciences. 2022. № 5. S. 294-300s.

10. Kontrol dodatkiv internet-trafika kompiuternykh merezh metodamy mashynnoho navchannia. / V.M. Dzhulii, Yu.P. Klots, I.V. Muliar, M.L. Zhylevych, A.V. Dzhulii // Herald of Khmelnytskyi National University. Technical sciences.– Khmelnytskyi. – 2021, – №5. – pp. 22–26.

11. Dzhulii, V.M. (), 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 KNU im/ Tarasa Shevchenka. – K.: VIKNU. – 2022. – Vyp. №74. – pp. 73-82.

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

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

14. Organizational Network Analysis as a Tool for Leadership Assessment in Software Development Team. / L.Yemchuk, O. Zhylinska; A. Chornyi; V. Dzhuliy // Institute of Electrical and Electronics Engineers (30 September 2020); INSPEC Accession №: 20008165; DOI: 10.1109/ACIT49673.2020.

ДОДАТОК В

Презентація кваліфікаційної роботи

Тема Метод захисту від витоку інформації на основі поділу стислих та зашифрованих даних

Мета магістерської роботи - полягає в підвищенні точності класифікації, сформованих алгоритмами стиснення та шифрування інформації, псевдовипадкових послідовностей.

Наукова задача – розробка методу класифікації, сформованих алгоритмами стиснення та шифрування інформації, псевдовипадкових послідовностей, для захисту від витоку конфіденційних даних в зашифрованому вигляді.

Об'єкт дослідження: Псевдовипадкові послідовності, сформовані алгоритмами стиснення та шифрування інформації.

Предмет дослідження: Алгоритми, методи класифікації, сформованих алгоритмами стиснення та шифрування інформації, псевдовипадкових послідовностей.

Задачі досліджень у роботі формулюються наступним чином:

1. 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.
2. Розробити модель, сформованих алгоритмами стиснення та шифрування даних, псевдовипадкових послідовностей, що відрізняється від відомих, врахуванням їх статистичних характеристик.
3. Розробити метод класифікації, сформованих алгоритмами стиснення та шифрування даних, псевдовипадкових послідовностей, що враховує здатність їх статистичних ознак.

Наукова новизна роботи визначає:

1. Модель, сформованих алгоритмами стиснення та шифрування інформації, псевдовипадкових послідовностей, відрізняється врахуванням статистичних характеристик послідовностей;

2. Метод класифікації псевдовипадкових послідовностей, сформованих алгоритмами шифрування та стиснення інформації, враховує статистичні ознаки послідовностей.

3. Спосіб класифікації, сформованих алгоритмами стиснення та шифрування інформації, псевдовипадкових послідовностей, для захисту від витоку конфіденційних даних в зашифрованому вигляді.

Методи дослідження. Для вирішення задач у магістерській роботі застосовувалися методи: теорії розпізнавання образів, математичної статистики, математичного моделювання.

Практична цінність Практична цінність магістерського дослідження полягає у підвищенні точності класифікації, сформованих алгоритмами стиснення та шифрування інформації, псевдовипадкових послідовностей, для захисту від витоку конфіденційних даних в зашифрованому представленні та відмові від контекстних ознак.

Апробація роботи. Наукові результати і основні положення магістерської роботи доповідались і обговорювались на всеукраїнських та міжнародних науково-технічних конференціях,

Публікації. За темою дипломної роботи ОКР «Магістр» опубліковано 1 теза доповідей, 1 фахова стаття.

Значення рівня показників небезпеки інформації

Показник рівня небезпеки інформації	Найменування показника небезпеки	Значення показників рівня небезпеки
Тип доступу (td)	Віддалений	3
	Фізичний	1
	Локальний	2
Рівень складності (dl)	Помірний	4
	Середній	2
	Підвищений	3
	Високий	1
Значимість інформаційних компонентів, ресурсів (rc)	Низька	1
	Середня	2
	Висока	3

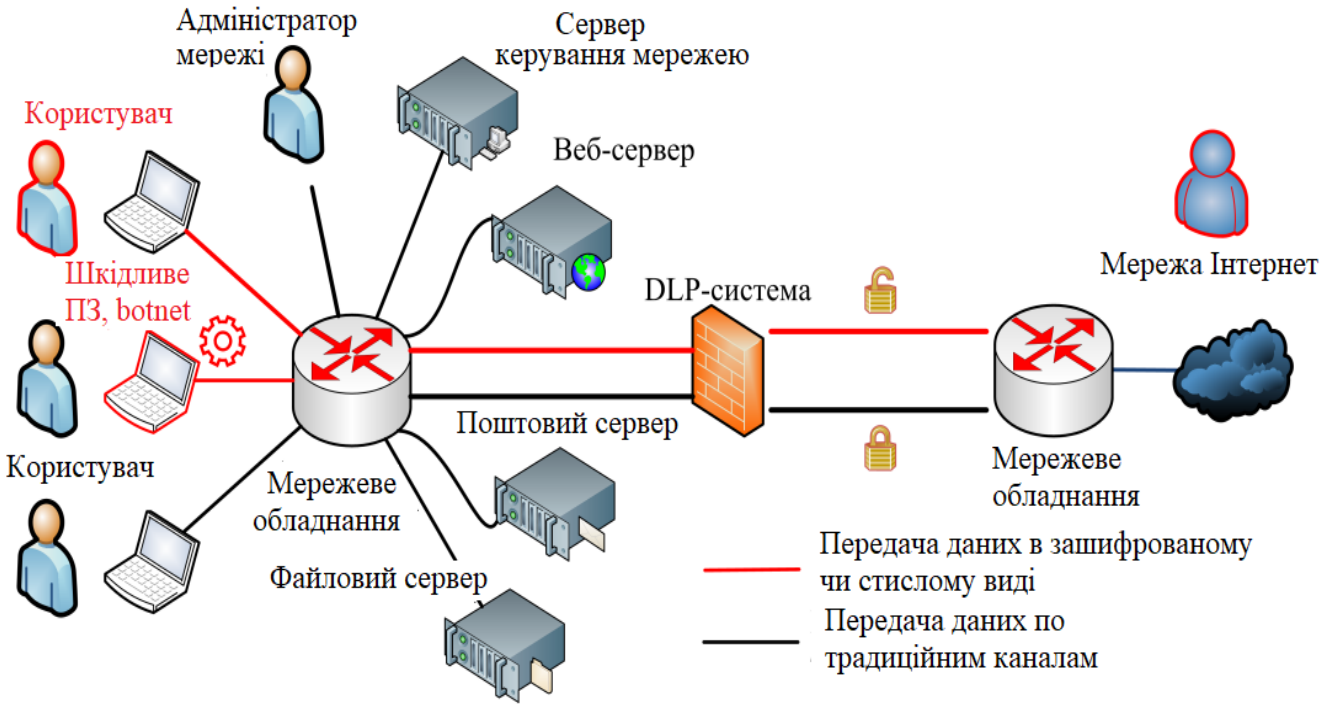
Показник рівня небезпеки загрози конфіденційним даним складається із суми значень показників і становить 8, відповідно до виразу (1):

$$(W) = td(2) + dl(4) + rc(2) = 8 \quad (1)$$

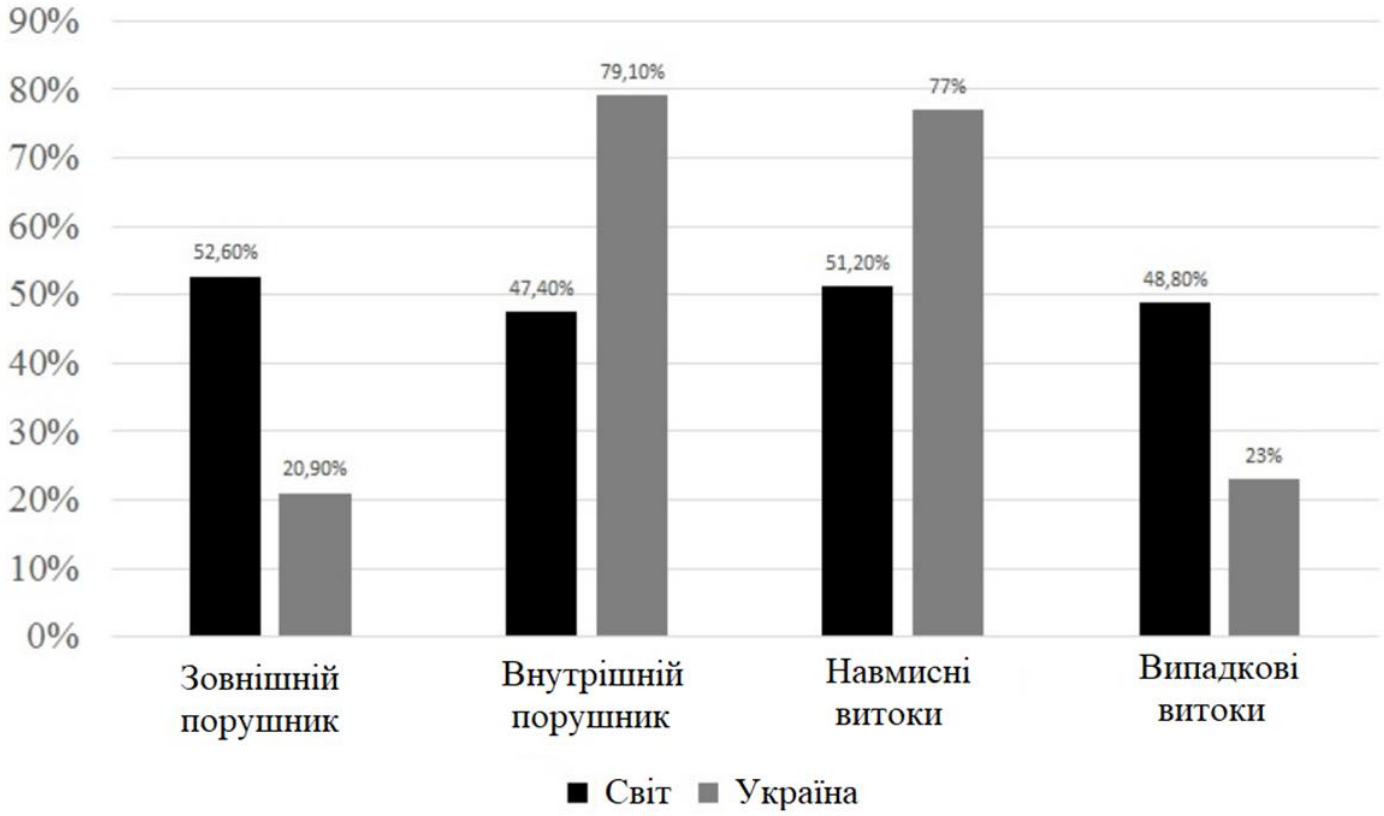
Рівень небезпеки загроз безпеки даних

Рівень загроз безпеки даних	Діапазон значень
Низький	$W = 4$
Середній	$5 \le W \le 7$
Високий	$W = 8$

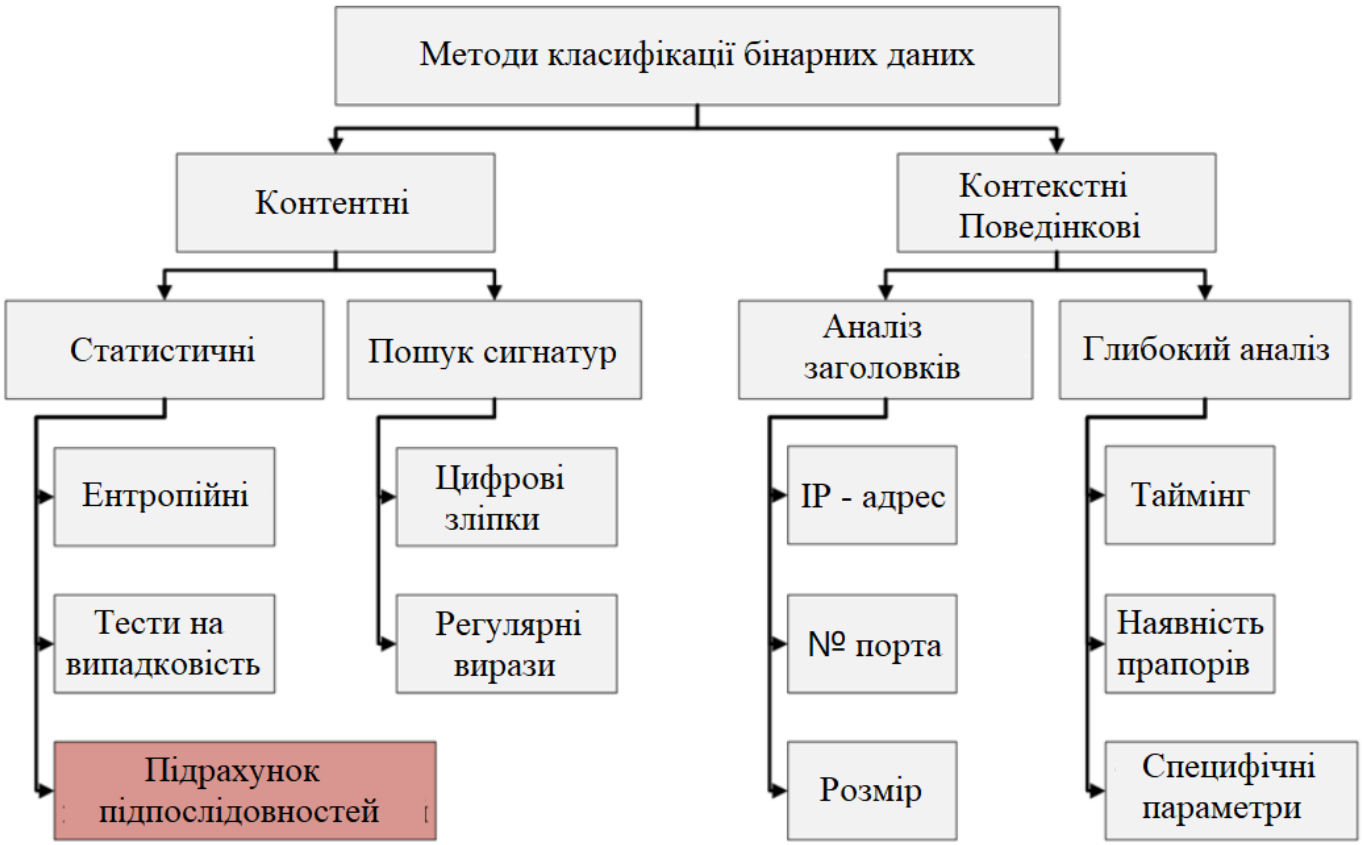
Схема процесу витоку даних, реалізована внутрішнім порушником



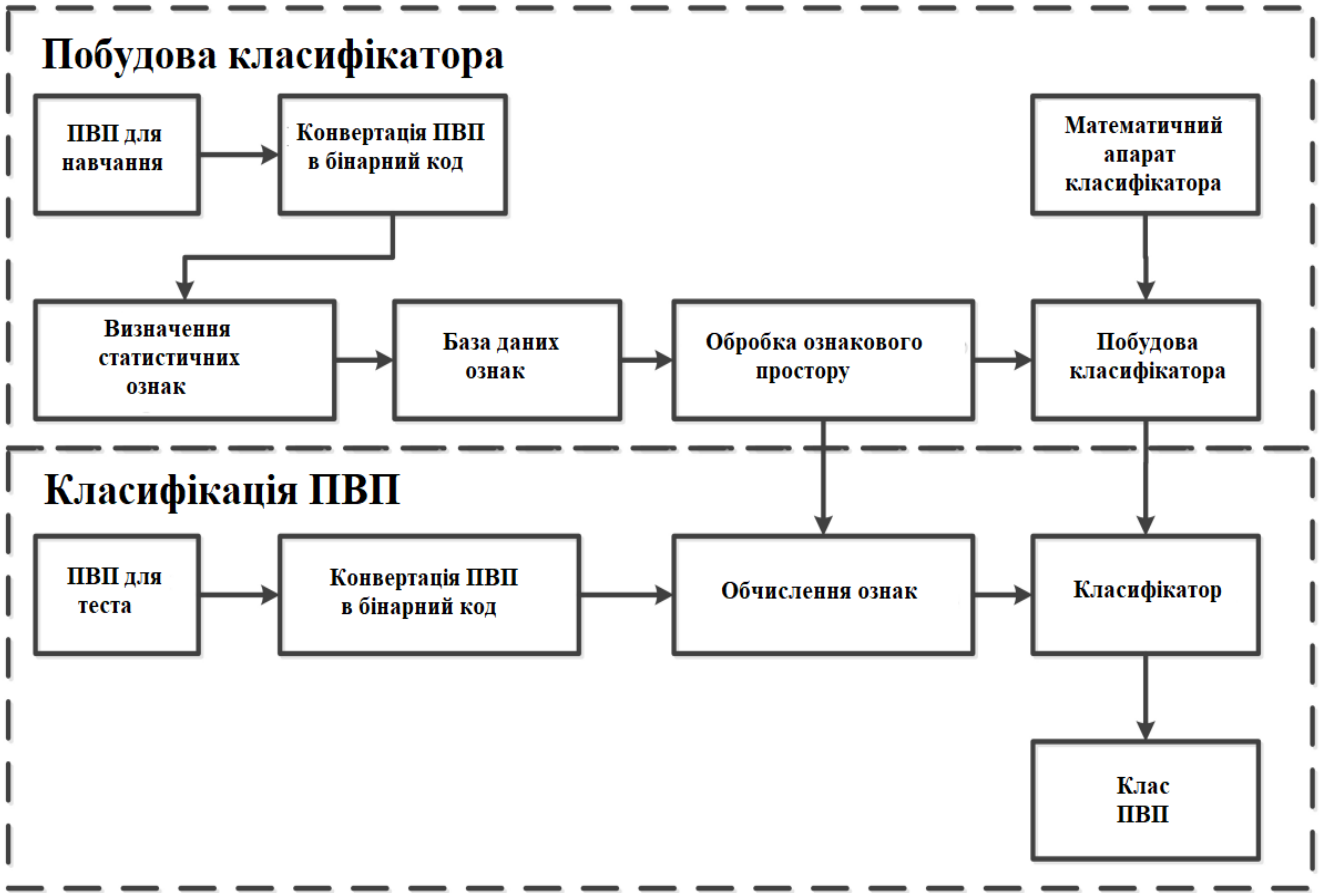
Статистика зафіксованих витоків інформації за 2021р.



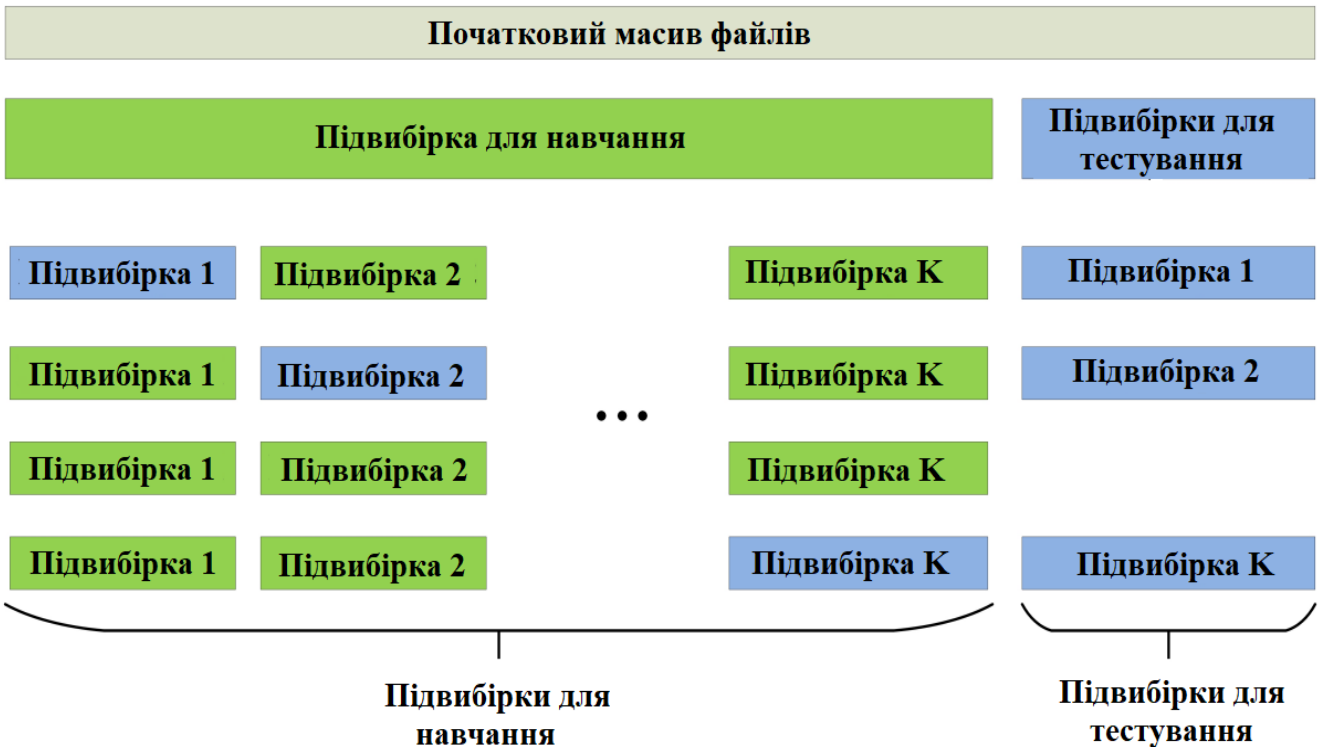
Класифікація методів аналізу інформації, що використовуються в DLP-системах



Функціональна модель процесу формування класифікатора



Процедура стратифікованого вибору груп



Модель псевдовипадкових послідовностей

Модель псевдовипадкових послідовностей - вектор статистичних характеристик, які обчислюються виразом (1):

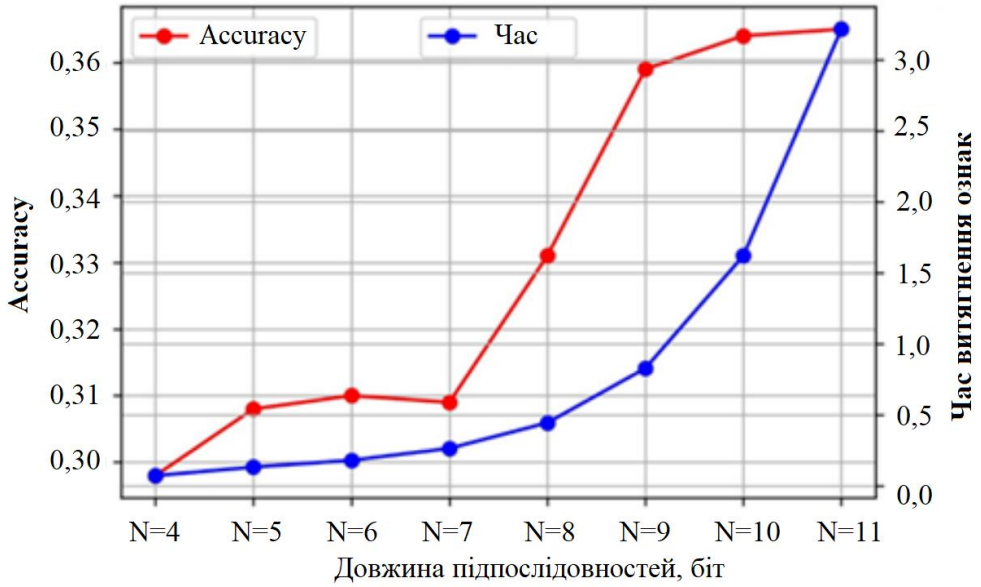
$$V_{Sub} = \left(f_j, ..., f_{2^N} \right) \tag{1}$$

Вхідна вибірка була перетворена на вісім наборів даних $V = V_4, ..., V_{11}$, містять вектори псевдовипадкових послідовностей, які визначаються виразом (2), і складаються зі значення частот підпослідовностей довжиною 4-11 біт.

$$f_j = F(j) = \frac{n(j)}{M - N(j) + 1}, j \in \left\{ 0, ..., 2^N \right\}, \tag{2}$$

де f_j - частота входження підпослідовності j в аналізовану послідовність, $n(j)$ - кількість входжень підпослідовності j в аналізовану псевдовипадкову послідовність, M – довжина аналізованої послідовності в бітах, $N(j)$ – довжина підпослідовності в бітах. Отримані ознакові простори подавалися на вхід алгоритму випадкового лісу для визначення точності класифікації псевдовипадкових послідовностей

Залежність точності класифікації ПВП на основі алгоритмому випадкового лісу



До моделі додані статистичні ознаки розподілу байт: середнє значення (B_{mean}), середньоквадратичне відхилення (B_{sko}), мінімальне ($b_{\min}$) та максимальні ($b_{\max}$) значення кількості байт у послідовності (3):

$$\begin{cases} B_{mean} = \dfrac{\sum_{i=0}^{255} n(b_i)}{256} \\ B_{sko} = \sqrt{\dfrac{\sum_{i=0}^{255} \left(n(b_i) - B_{mean} \right)^2}{256}} \\ b_{\min} = Min\left(n(b_i) \right) \\ b_{\max} = Max\left(n(b_i) \right) \end{cases} \tag{3}$$

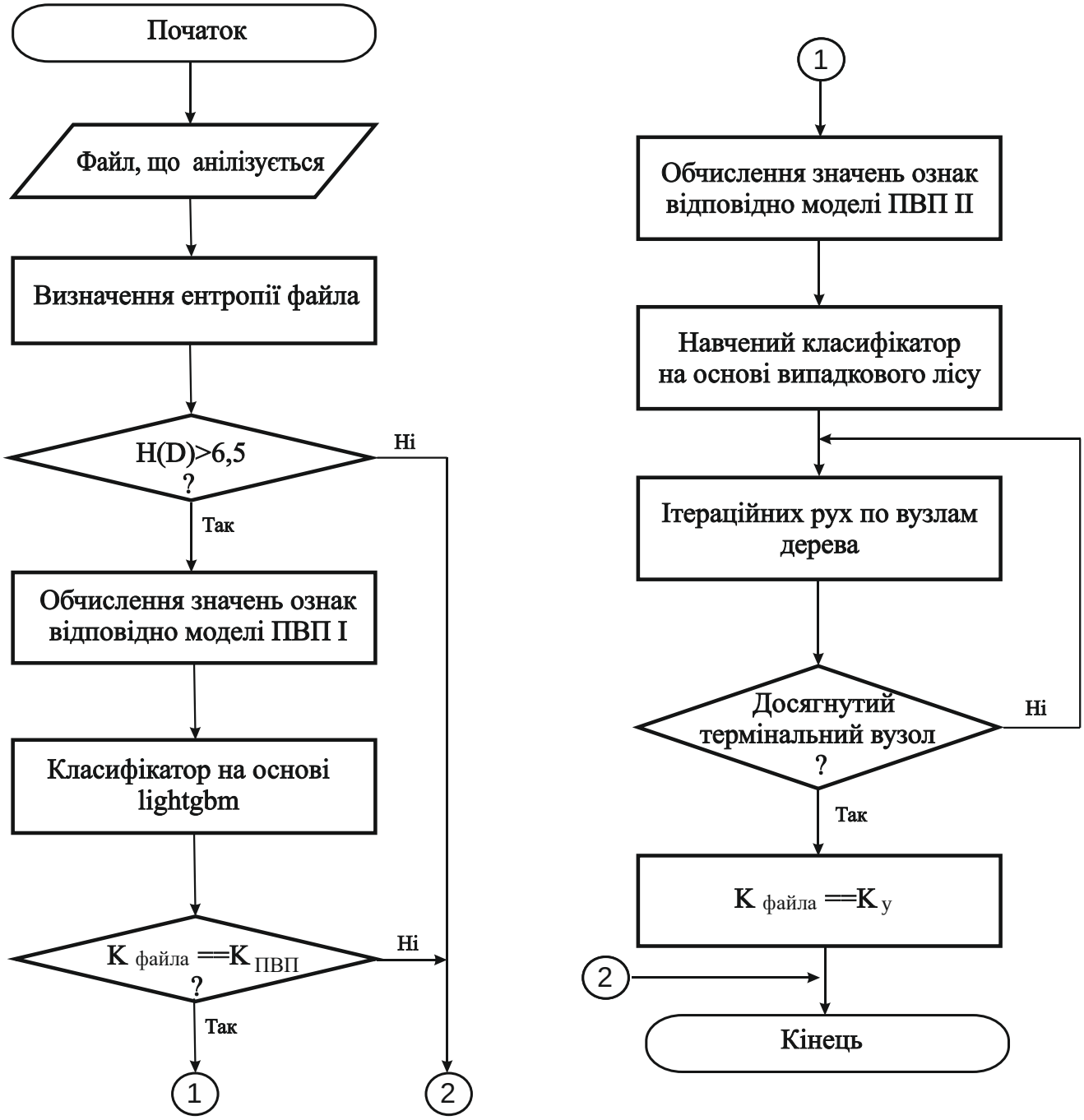
де $(n(b_i))$ – кількість появи i - байта в ПВП, яка піддається аналізу

Таким чином, модель псевдовипадкової послідовності задається виразом (4)

$$V_{stat} = \left\langle v_1, ..., v_\varphi \right\rangle = \left\langle f_j, ..., f_{2^N}, b_0, ..., b_{255}, B_{mean}, B_{sko}, b_{\min}, b_{\max} \right\rangle \tag{4}$$

Метод класифікації стиснених та зашифрованих ПВП

Алгоритм класифікації стислих та зашифрованих даних

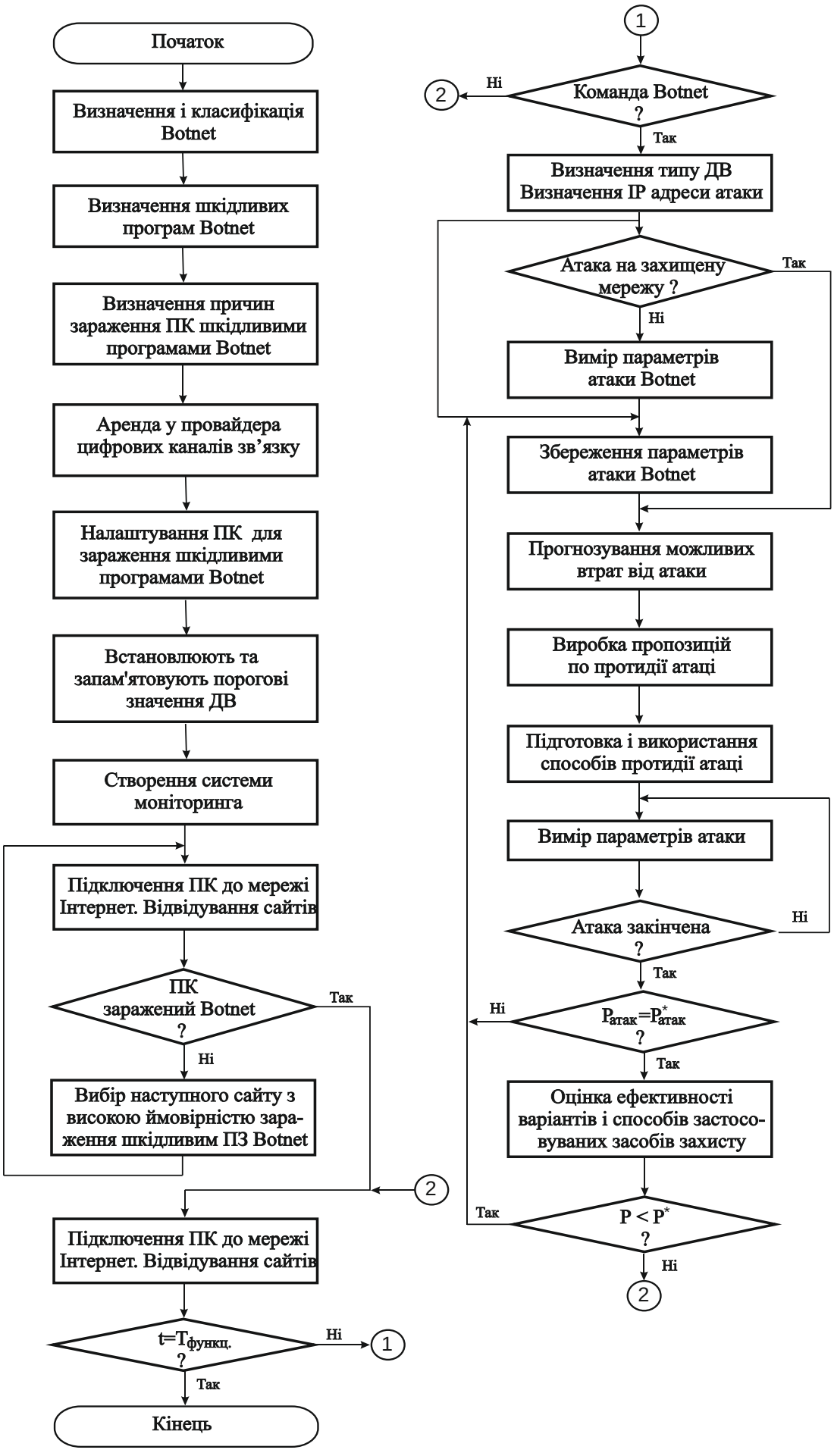


Редукування простору ознак виконується на основі обчислення локальних ваг. Критерієм розбиття визначається виразом (1)

$$\begin{cases} IG\left(D_p,\upsilon\right)=IG\left(D_p\right)-\dfrac{N_{left}}{N_p}\cdot I\left(D_{left}\right)-\dfrac{N_{right}}{N_p}\cdot I\left(D_{right}\right) \\ IG\rightarrow \max \end{cases} \quad (1)$$

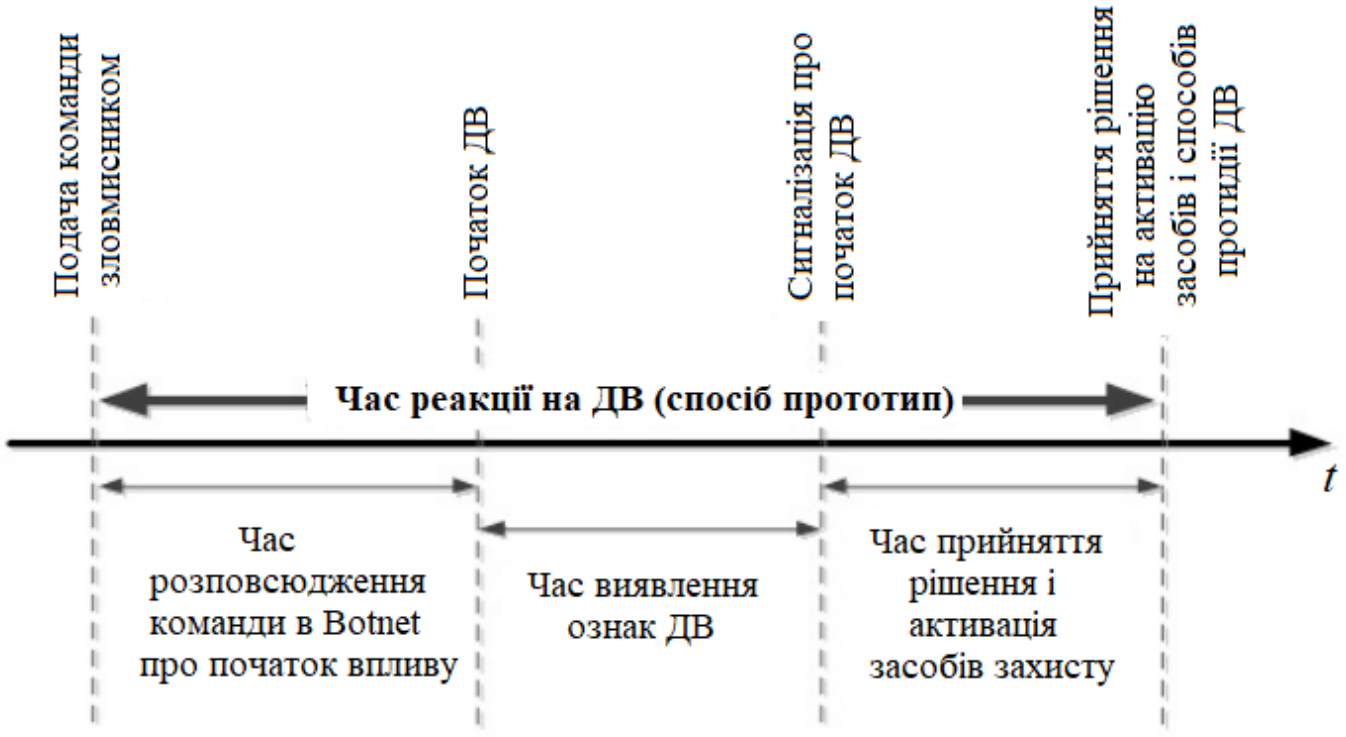
де $IG\left(D_p\right)$– приріст інформації після розподілу батьківського вузла, $I\left(D_p\right)$, $I\left(D_{left}\right), I\left(D_{right}\right)$– значення міри неоднорідності в батьківському вузлі, правому та лівому нащадках відповідно, $\upsilon\in V$ – ознака, за якою відбувається розбиття набору даних.

Алгоритм раннього виявлення атак Botnet на мережу передачі інформації

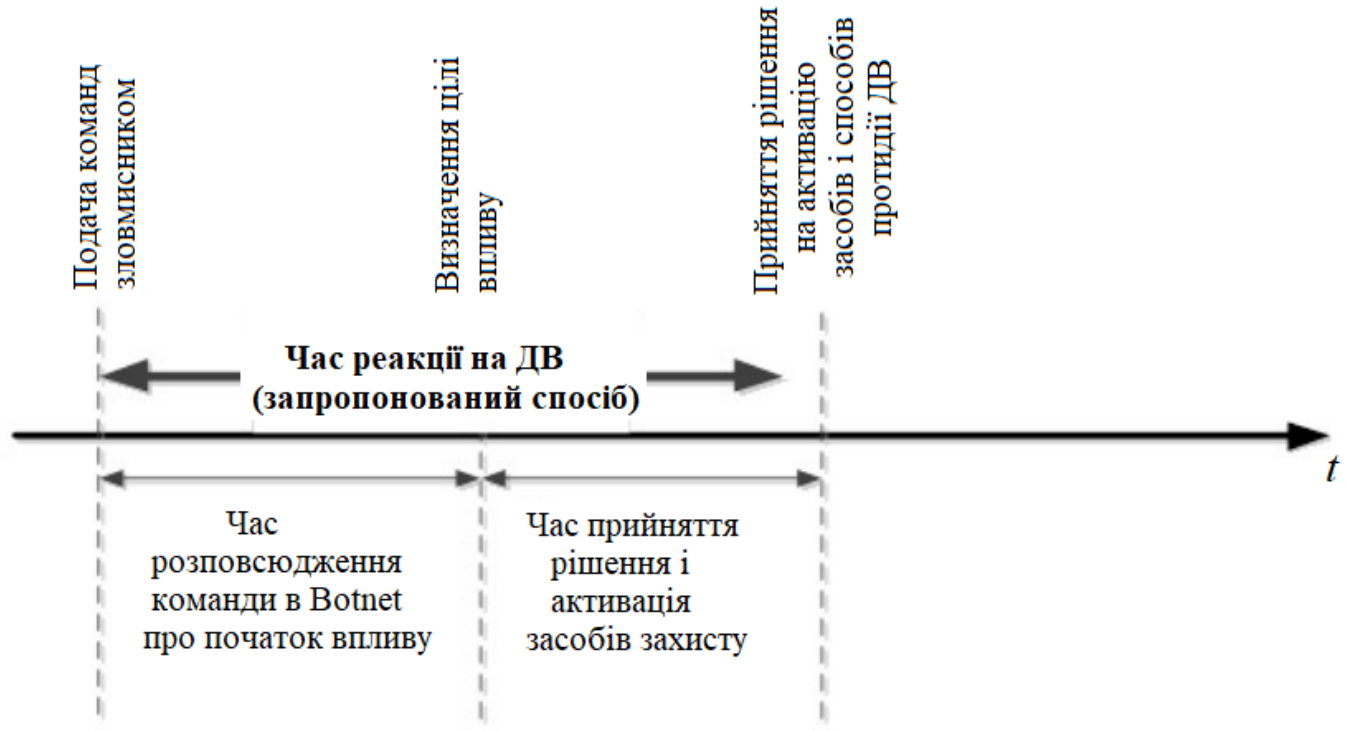


Вирішення проблеми раннього виявлення деструктивного впливу Botnet

(а) – спосіб прототип; б) -запропонований спосіб)



а)



б)

Схема впровадження запропонованого модуля статистичного аналізу даних у DLP-системи

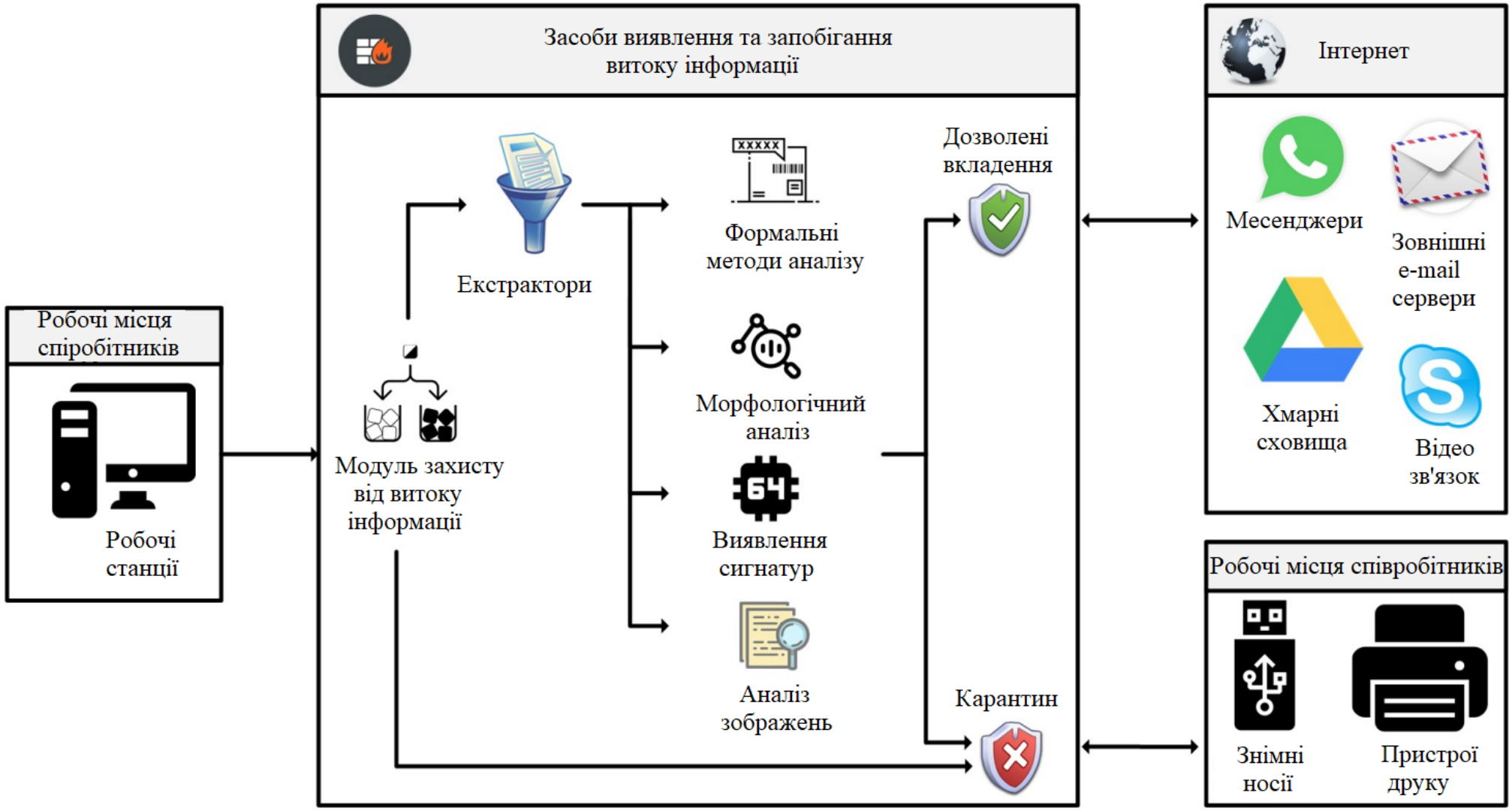
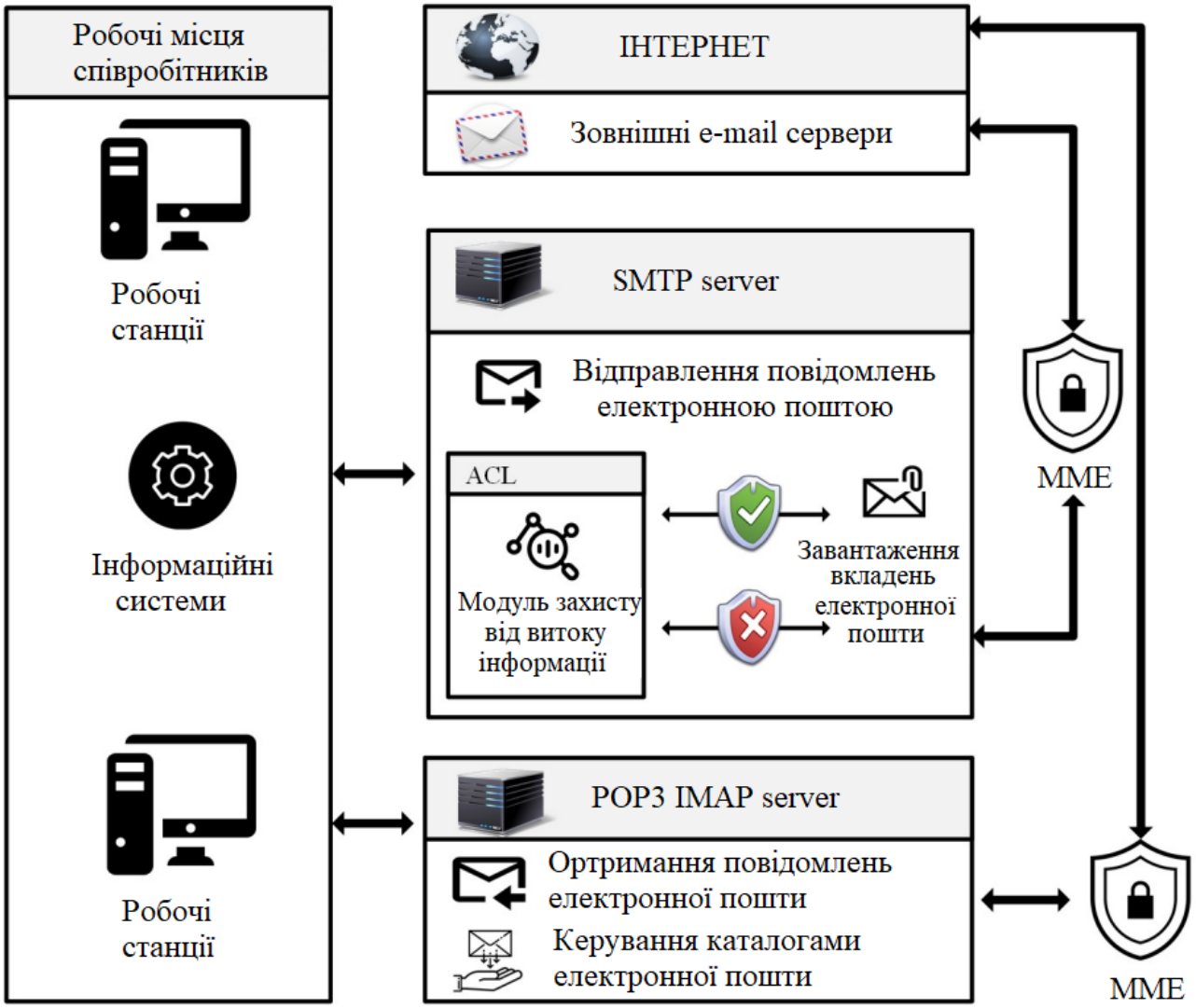
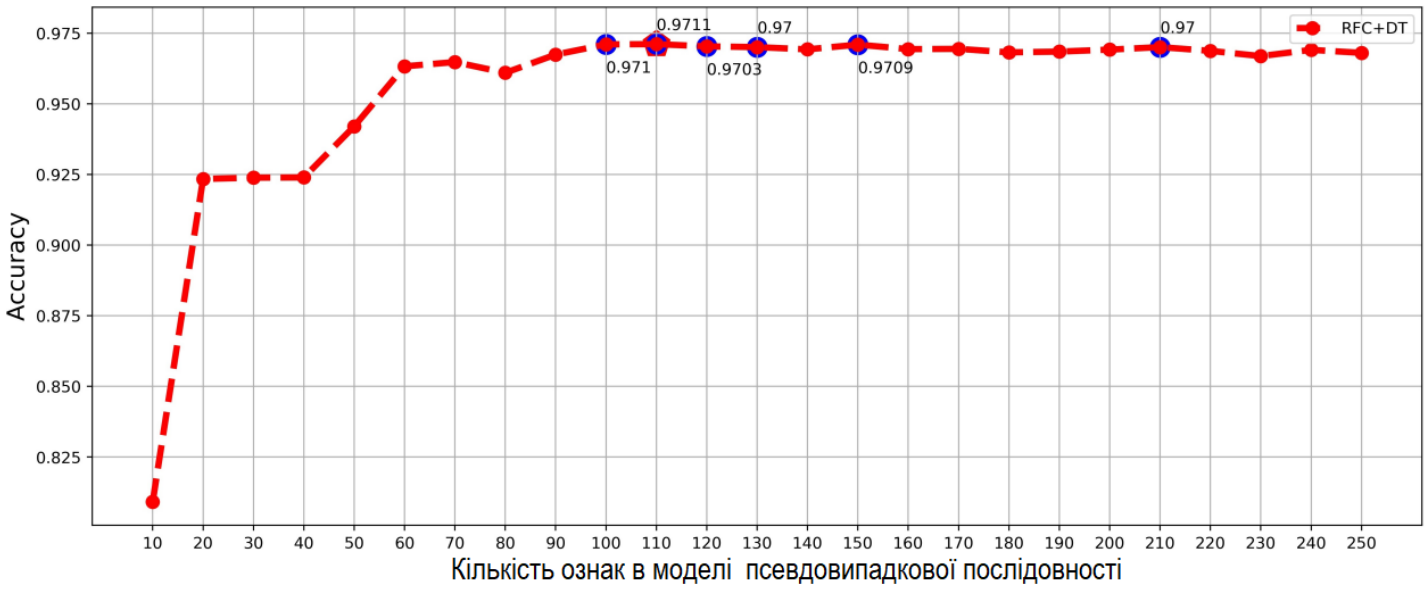


Схема впровадження запропонованого методу класифікації псевдовипадкових послідовностей на сервері пошти



Оцінка точності класифікатора від числа ознак моделі псевдовипадкової послідовності



ВИСНОВКИ

Основні результати магістерського дослідження полягають у наступному:

1. На основі проведеного аналізу особливостей функціонування відомих засобів запобігання та виявлення витоку інформації, виявлено обмеження підходів класифікації стиснених та зашифрованих даних, пов'язані з використання заголовків для проведення аналізу файлів та низькою точністю класифікації без використання заголовків.

2. У ході проведених експериментів визначено раціональну довжину підпослідовностей – дев’ять біт, що беруть участь у формуванні простору ознак.

3. Модель псевдовипадкових послідовностей, сформованих алгоритмами стиснення та шифрування даних, дозволила врахувати особливості стиснених та зашифрованих псевдовипадкових послідовностей при поданні в бінарному виді, підпослідовностями довжиною в дев’ять біт.

4. Метод класифікації псевдовипадкових послідовностей, сформованих алгоритмами стиснення та шифрування даних, враховує дискримінуючу здатність статистичних ознак послідовностей, показує більш високу точність класифікації на відміну від відомих аналогів.

5. Проведено оцінку ефективності запропонованих підходів. Отримані значення точності класифікації псевдовипадкових послідовностей перевищують відомі аналоги.

Отримані значення часу та точності класифікації запропонованого алгоритму в порівнянні з існуючими дослідженнями в заданій предметній області дозволяють зробити висновок про досягнення мети магістерського дослідження.

За темою роботи опубліковано 1 теза та 1 наукова стаття.

Завідувачу кафедри КБ
канд.техн.наук, доц. Кльоцу Ю. П.

Кучерявого Євгена Ігоровича
ПІБ здобувача вищої освіти

ФІТ, 2 курсу, групи КБ-22-1

ЗАЯВА

З правилами чинного Положення «Про систему забезпечення академічної доброчесності у Хмельницькому національному університеті» від 01.07.2022, згідно з яким виявлення плагіату є підставою для відмови в допуску кваліфікаційної роботи до захисту та застосування заходів дисциплінарної та академічної відповідальності, ознайомлений (а). Про використання програмно-технічних засобів для перевірки кваліфікаційних робіт здобувачів вищої освіти на плагіат оповіщений(а) та надаю свою згоду на обробку та збереження університетом моєї роботи в інституційному репозитарії університету.

Також надаю університету право на передачу моєї роботи для обробки та збереження в базах даних програмно-технічних засобів (Unicheck та Anti-Plagiarism) та використання роботи для виявлення плагіату в інших роботах, які перевіряються програмно-технічними засобами та користувачами, що мають доступ до цих програмно-технічних засобів, виключно в обмежених цілях для виявлення плагіату в текстах робіт.

Робота для перевірки університетом надається в друкованому та електронному варіанті. Електронна версія моєї роботи збігається (ідентична) з друкованою.

11 грудня 2023 року

Anti-Plagiarism v-15.257

Максимальне співпадіння з одним документом 0.0%

Словники перевірки: en_US, ru_RU, ua_UA. **Помилок в документах: 9%**

	Документ		Сумарний збіг по Базі Даних	
ID: 121898 Назва: Метод захисту від витоку інформації на основі поділу стислих та зашифрованих даних Додано в БД: 2023-12-06 Автора: Кучерявий Є.І. Керівники: Джулій В.М. Консультанти: Опоненти:	Символи	Лексеми	Символи	Лексеми
	105563	708	1067 (1%)	17 (2%)

Джерело плагіату

ID	Опис	Наявність плагіату в документі	
		Символи	Лексеми

UNICHECK
by Turnitin

Ім'я користувача:
Кафедра кібербезпеки

ID перевірки:
1015975312

Дата перевірки:
06.12.2023 10:35:26 EET

Тип перевірки:
Doc vs Internet + Library

Дата звіту:
06.12.2023 10:46:04 EET

ID користувача:
100008300

Назва документа: **Кучерявий_Магістерська_Плагіат**

Кількість сторінок: **78** Кількість слів: **13573** Кількість символів: **112765** Розмір файлу: **12.73 MB** ID файлу: **1015654799**

Виявлено модифікації тексту (можуть впливати на відсоток схожості)

1.61%
Схожість

Найбільша схожість: **0.77%** з джерелом з Бібліотеки (ID файлу: **1015654797**)

1.41% Джерела з Інтернету	177	Сторінка 80
1.07% Джерела з Бібліотеки	69	Сторінка 81

0% Цитат

Вилучення цитат вимкнене

Вилучення списку бібліографічних посилань вимкнене

0%
Вилучень

Немає вилучених джерел

Модифікації

Виявлено модифікації тексту. Детальна інформація доступна в онлайн-звіті.

Замінені символи	29
Підозріле форматування	12 сторінок

РІШЕННЯ ЕКСПЕРНОЇ КОМІСІЇ

КАФЕДРИ КІБЕРБЕЗПЕКИ

ПРО ДОПУСК КВАЛІФІКАЦІЙНОЇ РОБОТИ ДО ЗАХИСТУ

Підтверджуємо ознайомлення з результатом звіту подібності щодо роботи, генерованого системою виявлення текстових збігів/ідентичності/схожості:

Назва: Метод захисту від витоку інформації на основі поділу стислих та зашифрованих даних

Автор: Євген КУЧЕРЯВИЙ
Спеціальність: 125 – Кібербезпека
Освітня програма: Кібербезпека
Науковий керівник: Володимир ДЖУЛІЙ., к.т.н, доц.
Після аналізу звіту подібності зроблено такий висновок:

№	Висновок	Позначка про відповідність
1	Запозичення, виявлені в роботі, є законними і не є плагіатом (далі – зазначаються підстави віднесення запозичень до правомірних). Робота приймається до захисту.	відповідає
2	Виявлені запозичення не є плагіатом, розміщені в розділах, які не описують безпосередньо авторське дослідження, але кількість цитат перевищує обсяг, виправданий поставленою метою роботи (далі – зазначаються детальні та аргументовані підстави віднесення запозичень до правомірних). Робота приймається до захисту, але має бути відкоригована. Відкоригований варіант має бути поданий на кафедру за 2 дні до захисту, разом із заявою щодо самостійності виконання письмової роботи та ідентичності друкованої та електронної версії роботи.	
3	Виявлені запозичення не є плагіатом, але частково розміщені в розділах, які описують безпосередньо авторське дослідження, а кількість цитат перевищує обсяг, виправданий поставленою метою роботи. В зв'язку з цим мета роботи та поставлені завдання не були досягнені. Робота може бути допущена до захисту (наступного року) після того як буде відкоригована та допрацьована і успішно пройде повторну перевірку на академічний плагіат.	
4	Робота містить навмисні текстові спотворення, передбачувані спроби укриття запозичень або інші прояви академічного плагіату. Робота містить фабрикацію або фальсифікацію даних. Робота не допускається до захисту.	
5	Інше:	

Підтвердження:

Запозичення, виявлені в роботі, є законними і не є плагіатом, оскільки:

1) запозичення розміщені в розділах аналізу існуючих аналогів та прототипів, які не описують безпосередньо авторське дослідження і не стосуються результатів роботи;
2) усі запозичення фрагментарні, або мають належним чином оформленні посилання;
3) всі зафіксовані системою ознаки модифікації тексту відносяться до комбінування латинських символів зі україномовними скороченнями індексів в формулах, що не є модифікацією тексту.

Сумарний обсяг всіх запозичень, визначений системою виявлення збігів/ ідентичності/схожості, складає 1.61% і адресується до 177 першоджерела, що, з урахуванням наведених обґрунтувань, відповідає характеру наукового дослідження і свідчить на користь кваліфікаційної роботи.

Керівник роботи _____ Володимир ДЖУЛІЙ

Завідувач кафедри кібербезпеки _____ Юрій КЛЬОЦ

Дата: 11.12.2023

гарант ОП Віра Тітова

МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ
ХМЕЛЬНИЦЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ

РЕЦЕНЗІЯ НА КВАЛІФІКАЦІЙНУ РОБОТУ

освітньо-кваліфікаційного рівня «магістр»

Студент Кучерявий Євген Ігорович

Тема: «Метод захисту від витоку інформації на основі поділу стислих та зашифрованих даних»

Галузь знань 12 «*Інформаційні технології*» Спеціальність 125 «*Кібербезпека*» Освітня програма «Кібербезпека»

Обсяг кваліфікаційної роботи освітньо-кваліфікаційного рівня «магістр»: кількість сторінок записки 82;

1. Короткий зміст КР та прийнятих рішень Кваліфікаційна робота присвячена дослідженню питань, пов'язаних з розробкою та впровадженням системи захисту від витоку інформації на основі поділу стислих та зашифрованих даних, з метою підвищення точності класифікації, сформованих алгоритмами стиснення та шифрування інформації, псевдовипадкових послідовностей. Для досягнення мети проведено дослідження особливостей функціонування перспективних засобів запобігання та виявлення витоку конфіденційних даних, виявлено обмеження, пов'язані з виявленням стислої та зашифрованої інформації, обґрунтувано вибір відповідного ознакового простору для моделювання, сформованих алгоритмами стиснення та шифрування інформації, псевдовипадкових послідовностей; розроблено модель псевдовипадкових послідовностей, що відрізняється від відомих, врахуванням їх статистичних характеристик, розроблено метод класифікації псевдовипадкових послідовностей, що враховує здатність їх статистичних ознак
2. Висновок про відповідність КР завданню Кваліфікаційна робота у повній мірі відповідає поставленому завданню як в теоретичній так і у практичній частині роботи.
3. Характеристика виконання кожного розділу роботи, ступінь використання останніх досягнень науки і техніки і передових методів роботи: У вступі обґрунтовується актуальність теми роботи, її зв'язок з галуззю знань «Інформаційні технології» та спеціальністю «Кібербезпека», формулюється мета та основні завдання кваліфікаційної роботи. У першому розділі проведено аналіз існуючих систем захисту та управління доступом, що дозволило виявити проблеми та завдання, що потребують вирішення; застосування принципів побудови систем захисту від витоку інформації на основі поділу стислих та зашифрованих даних стало основою для постановки задачі і проєктування архітектури системи. У другому розділі проаналізовано вимоги та потреби підприємства з урахуванням специфіки, визначено необхідні компоненти та функціонал системи захисту від витоку інформації на основі поділу стислих та зашифрованих даних, побудована модель системи захисту від витоку інформації. У третьому розділі наведено опис процесу реалізації системи системи захисту від витоку інформації на підприємстві.

4. Позитивні сторони кваліфікаційної роботи Модель псевдовипадкових послідовностей, сформованих алгоритмами стиснення та шифрування даних, дозволила врахувати особливості стиснених та зашифрованих псевдовипадкових послідовностей при поданні в бінарному виді, підпослідовностями довжиною в дев'ять біт. Метод класифікації псевдовипадкових послідовностей, сформованих алгоритмами стиснення та шифрування даних, враховує дискримінуючу здатність статистичних ознак послідовностей, показує більш високу точність класифікації на відміну від відомих аналогів. Проведено оцінку ефективності запропонованих підходів. Отримані значення точності класифікації псевдовипадкових послідовностей перевищують відомі аналоги.

5. Негативні сторони проекту: При застосування алгоритма класифікації, яким чином обчислюються локальні ваги та як при цьому визначається критерій розбиття ознак

6. Оцінка графічного оформлення та пояснювальної записки роботи.

7. Відгук про роботу в цілому В загальному кваліфікаційна робота заслуговує позитивної оцінки. Весь матеріал кваліфікаційної роботи структурований, чіткий та послідовний. Усі розділи роботи послідовні та логічні, що дозволяє чітко розуміти викладений матеріал в рамках тематики кваліфікаційної роботи. У пояснювальній записці багато наглядних пояснень.

8. Інші зауваження -

9. Оцінка дипломної роботи Розглянувши позитивні та негативні сторони представленої кваліфікаційної роботи, можна зробити висновок, що робота заслуговує оцінки «добре/ С (4,0)».

РЕЦЕНЗЕНТ (прізвище, ім'я, по батькові, посада, місце роботи)

Лисенко Сергій Миколайович, д.т.н., професор, кафедра комп'ютерної інженерії та інформаційних систем, Хмельницького національного університету

« 11 » грудня 2023 . (підпис)