УДК 004.891

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

DOI:https://doi.org/10.17721/2519-481X/2024/82-09

МЕТОД КЛАСИФІКАЦІЇ ПСЕВДОВИПАДКОВИХ ПОСЛІДОВНОСТЕЙ СТИСЛИХ ТА ЗАШИФРОВАНИХ ДАНИХ ЗАПОБІГАННЮ ВИТОКУ ІНФОРМАЦІЇ

Розглянута задача розробки методу класифікації псевдовипадкових послідовностей захисту від витоку конфіденційної інформації на основі поділу стислих та зашифрованих даних, може бути використана для виявлення мережевих атак на мережі передачі даних, в засобах запобігання та виявлення витоку інформації, а також в програмних продуктах, що реалізують сервіси електронної пошти.

Показано, що загрози безпеки даних характеризуються набором векторних показників, кількісних та якісних, для їх формалізації необхідне застосування теорії нечітких множин та дискретної математики. Показано неможливість застосування експертних традиційних методів оцінок для визначення більшості розглянутих показників. Для мінімізації ризику витоку конфіденційної інформації пропонується формувати групи співробітників та розраховувати ризик витоку конфіденційних даних для кожної з них.

Сучасні засоби запобігання та виявлення витокам інформації застосовують різні методи проведення аналізу потоку даних. До основних відносяться контекстні та контентні методи. Наведені методи не здатні виявити витік даних у стислому та зашифрованому виді, а додавання цифрових сигнатур дозволяє маскувати зашифровані дані під стислі простим способом, в області безпеки інформації знайшли широке використання поведінкові методи аналізу потоку даних та алгоритми машинного навчання. Однією з основних складностей, в даній ситуації, виступає побудова моделей даних, обробка та пошук ознакового простору.

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. Зашифрований потік даних може передаватися з робочих станцій співробітників, різних інформаційних систем, мережевих сховищ.

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.

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 раціональне, що дозволяє досягти класифікації псевдовипадкових послідовностей високої точності та мінімального часу виконання процедури класифікації. Обґрунтовано вибір скануючого оптимального вікна класифікатора розміром в 500 кб. Залежно від вимог до точності та швидкості аналізу даних запропоновано два режими роботи: сканування випадково вибраного фрагмента файла розміром 500 кб; сканування всього файлу скануючим вікном розміром 500 кб.

Наведено опис місць впровадження запропонованого методу 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.

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

Вступ. 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 інформації, зростає кількість витоків конфіденційної інформації. Однією з головних причин витоків конфіденційної інформації - наявність внутрішнього порушника, здатного дотримуватись встановлених правил та заходів роботи з даними, здійснювати передачу конфіденційної інформації за контрольований інформаційний периметр компанії [1 – 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 [5, 15 – 23].

Основними механізмами захисту від витоків даних, є методи, засновані на пошуку регулярних виразів, сигнатур, цифрових зліпків, виявлення аномалій, застосування алгоритмів машинного навчання [13,16, 24 – 26].

У сучасній компанії бізнес процеси неможливі без застосування корпоративних мереж передачі даних та перспективних інформаційних систем. Якщо механізми захисту даних від зовнішніх загроз досягли відповідних гарантованих рівнів, то способи та методи протидії інсайдеру слабо розвинені, в більшості документів, що регламентують політику безпеки кофіденційним даним компанії, містяться постулати про відсутність інсайдера, що тягне до зростання ймовірності порушення інформаційної безпеки конфіденційних даних [6,10,14,18].

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

Загрози безпеки конфіденційних даних характеризуються набором векторних показників, кількісних та якісних, для їх формалізації необхідне застосування теорії нечітких множин та дискретної математики. Показано неможливість застосування експертних традиційних методів оцінок для визначення більшості розглянутих показників. Проведено аналіз байєсовського підходу вирішення задачі, доведено, необхідність проведення аналізу великої кількості статистичних даних [10,14,18,19].

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 [7,8,19,20,23].

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 [8-12,18].

Для мінімізації ризику витоку конфіденційної інформації пропонується формувати групи співробітників та розраховувати ризик витоку конфіденційних даних для кожної з них. Запропонований підхід передбачає використання data leakage prevention (DLP) та security information and event managment (SIEM) систем. Причиною витоку даних можуть бути політичні, індивідуальні, фінансові мотиви працівників компанії [5,6,14,19,20].

Аналіз останніх досліджень та постановка задачі. Аналіз досліджень мережевої активності корпоративної мережі є ключовим компонентом запобігання та раннього виявлення загроз та атак безпеки конфіденційним даним, що виходять від інсайдерів. Логування подій безпеки та функціонування інформаційної системи можуть використовуватись у реальному часі для проведення аналізу, проте записи необхідно відфільтровувати, оскільки не всі дозволяють виявити загрозу, атаку безпеки даних.

Проведений аналіз відкритого та зашифрованого трафіку на основі підрахунку ентропії окремих слів довжиною 2..64 біт, потоку даних, стандартного відхилення та середнього значення зазначених величин. Для проведення експериментів використані пакети довжиною понад 20 байт: відкритих та зашифрованих. Найкращі результати досягнуті при використання алгоритму C5.0, точність класифікації - 0,978.

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

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

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 [8-11,15,19,20].

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 держави. Атаки на такі критичні інформаційні системи включають проникнення в мережу організації та встановлення шкідливого програмного забезпечення, які можуть розкрити конфіденційну інформацію, змінити поведінку технічного обладнання [14,23,25].

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 [7,14,19,20].

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 метод, сформований алгоритмами стиснення та шифрування даних, псевдовипадкових послідовностей, що відрізняється від відомих, врахуванням їх статистичних характеристик.

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.

Метод класифікації псевдовипадкових послідовностей стислих та зашифрованих даних. Метод класифікації псевдовипадкових послідовностей стислих та зашифрованих даних дозволяє підвищити точність класифікації при використанні моделі псевдовипадкових послідовностей за рахунок застосування комплексу алгоритмів з навчання класифікатора, виявлення найбільш значущих дискримінуючих ознак [7].

Вхідними даними для побудови класифікатора псевдовипадкових послідовностекй - множина стислих і зашифрованих послідовностей (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: швидкий, шляхом перевірки випадково вибраного вікна розміром 500 кбайт, повна перевірка файлу скануючим вікном розміром 500 кбайт.
4. Обчислення значень ознак згідно з моделлю псевдовипадкових послідовностей [7].
5. Визначення типу файла навченим класифікатором на основі градієнтного бустингу.
6. Якщо визначено клас послідовностей, то продовжити роботу алгоритму, інакше закінчити.
7. Обчислення значень ознак згідно з моделлю псевдовипадкових послідовностей.
8. Передача отриманих значень ознак на навчений класифікатор.
9. Ітераційний рух вузлами дерева випадкового лісу.
10. Визначення досягнення термінального вузла.
11. Визначення класу псевдовипадкової послідовності.

Алгоритм класифікації псевдовипадкових чисел представлений на рис. 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% вихідних послідовностей, решта 20% відносять до тестової вибірки. Для забезпечення незалежності методу захисту від витоку конфіденційної інформації виконується відкидання перших десять кілобайт файла, що аналізується та підрахунок ентропії.

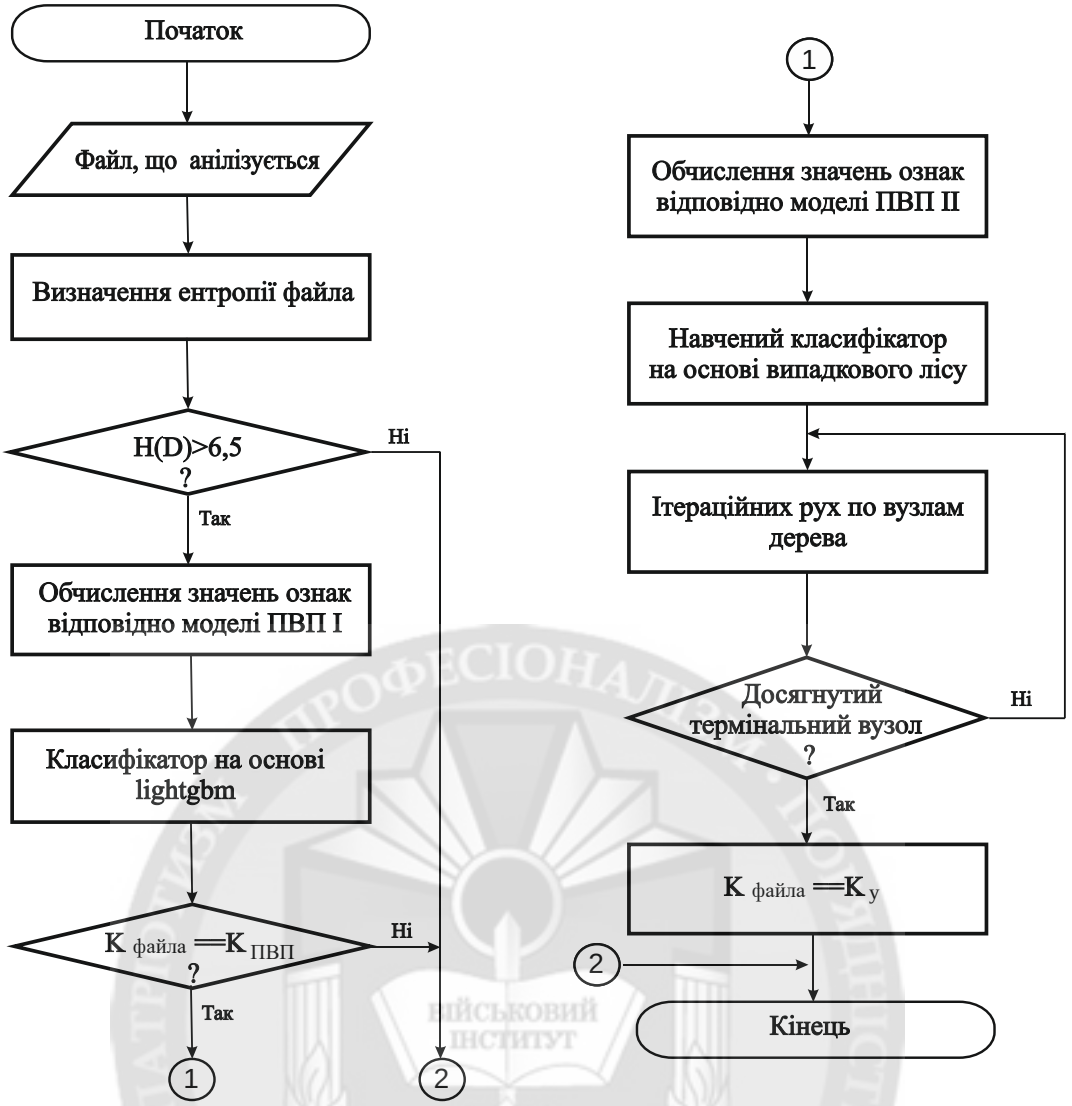

Рисунок 1 - Алгоритм класифікації стислих та зашифрованих даних

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

На третьому етапі, роботи алгоритму, відбувається формування ознакового простору та обчислення значень статистичних ознак згідно з виразом 1.

$$V_{stat} = \left\langle v_1, ..., v_\varphi \right\rangle = \left\langle f_j, ..., f_{2^N}, b_0, ..., b_{255}, B_{mean}, B_{sko}, b_{\min}, b_{\max} \right\rangle \quad (1)$$

Ознаковий простір складається з 512 значень частот входження підпослідовностей довжиною дев'ять біт $\left\langle f_j \right\rangle$, $j \in [0, ..., 511]$, 256 частот входження байт $\left\langle b_i \right\rangle$, $i \in [0, ..., 255]$ та чотири статистичних характеристик $\left\langle B_{mean}, B_{sko}, b_{\min}, b_{\max} \right\rangle$. 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, роботи алгоритму, відбувається формування редукованої множини ознакового простору. Розмірності простору здійснюється за рахунок випадково обраних ознак, дають мінімальне зменшення градієнта та використання ознак, що надає максимальне

зниження градієнта функції втрат. Вибір просторових ознак з мінімальним градієнтом дозволяє зберегти точність класифікації, знизити ознаковий простір.

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 аналіз, дозволяє аналізувати випадково обраний фрагмент розміром 500 кбайт; послідовний, полягає в проходженні всього файлу скануючим вікном розміром 500 кбайт.

На шостому етапі здійснюється обчислення ознак, редукованих за допомогою навченого класифікатора на основі випадкового лісу. Редукування простору ознак виконується на основі обчислення локальних ваг. Критерій розбиття визначається виразом 2.

$$\begin{cases} IG\left(D_p,\upsilon\right)=IG\left(D_p\right)-\dfrac{N_{left}}{N_p}\cdot I\left(D_{left}\right)-\dfrac{N_{right}}{N_p}\cdot I\left(D_{right}\right), \\ IG\to\max \end{cases} \tag{2}$$

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 виразом 3.

$$\begin{cases} Gini=1-\sum\limits_{y=1} p_y^2 \\ Entropy=-\sum\limits_{y=1} p_y\cdot\log_2 p_y \\ Error=1-\max\left(p_y\right) \end{cases}, \tag{3}$$

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.

$$w_{\upsilon_\varphi}^{\psi}=F\left(n_{\upsilon_\varphi}^{\psi}\right), \tag{4}$$

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 $[0{,}1\ldots 0{,}9]$, має виконуватися умова 5.

$$\begin{cases} \alpha+\beta+\gamma=1 \\ 0\le\alpha,\beta,\gamma\le 1 \end{cases} \tag{5}$$

Далі здійснюється визначення глобальних ваг відповідно до виразу 6.

$$\begin{cases} W_{\upsilon_\varphi} = F\left(\sum\limits_{\psi=1}^{\Psi} w_{\upsilon_\varphi}^{\psi}\right) \\ W_{\upsilon_\varphi}^{Global} = \min\limits_{\alpha,\beta,\gamma\in R} F\left(\upsilon_\varphi,\psi\right) = \alpha\cdot Entropy\left(\psi, w_{\upsilon_\varphi}^{\psi}\right) + \\ \quad +\beta\cdot Gini\left(\psi, w_{\upsilon_\varphi}^{\psi}\right) + \gamma\cdot Error\left(\psi, w_{\upsilon_\varphi}^{\psi}\right) \end{cases} \quad (6)$$

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. У разі присвоєння інформації мітки зашифрованих (стислих) даних, файл поміщається в карантин і відбувається генерація події безпеки даних відповідно до прийнятої політики в організації.

Алгоритм витягнення 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. 2.

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: середньоквадратичне відхилення, математичне очікування, максимальне та мінімальне значення частот входження байт. Результат роботи алгоритму формується в кортеж.

Реалізація методу класифікації псевдовипадкових послідовностей стислих та зашифрованих даних. Підсистеми статистичного аналізу інформації на основі методу класифікації псевдовипадкових послідовностей реалізована з використанням мови програмування Python, реалізує метод захисту від витоку інформації на основі поділу стислих та зашифрованих даних, може бути використана для виявлення мережевих атак на мережі передачі даних, в існуючих засобах запобігання та виявлення витоку інформації, а також в програмних продуктах, 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 (ДВ) на мережу, формується база даних про параметри деструктивних впливів на вузли мережі. Під час моніторингу деструктивних впливів 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.

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 2 - Алгоритм витягнення ознак із псевдовипадкових послідовностей

На основі отриманих статистичних даних визначають шкідливе програмне забезпечення, з використанням яких зловмисник заражає персональний комп'ютер (ПК) і команди, що надсилаються центром управління Botnet, зараженим ПК. 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 надійшли, 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. 3.

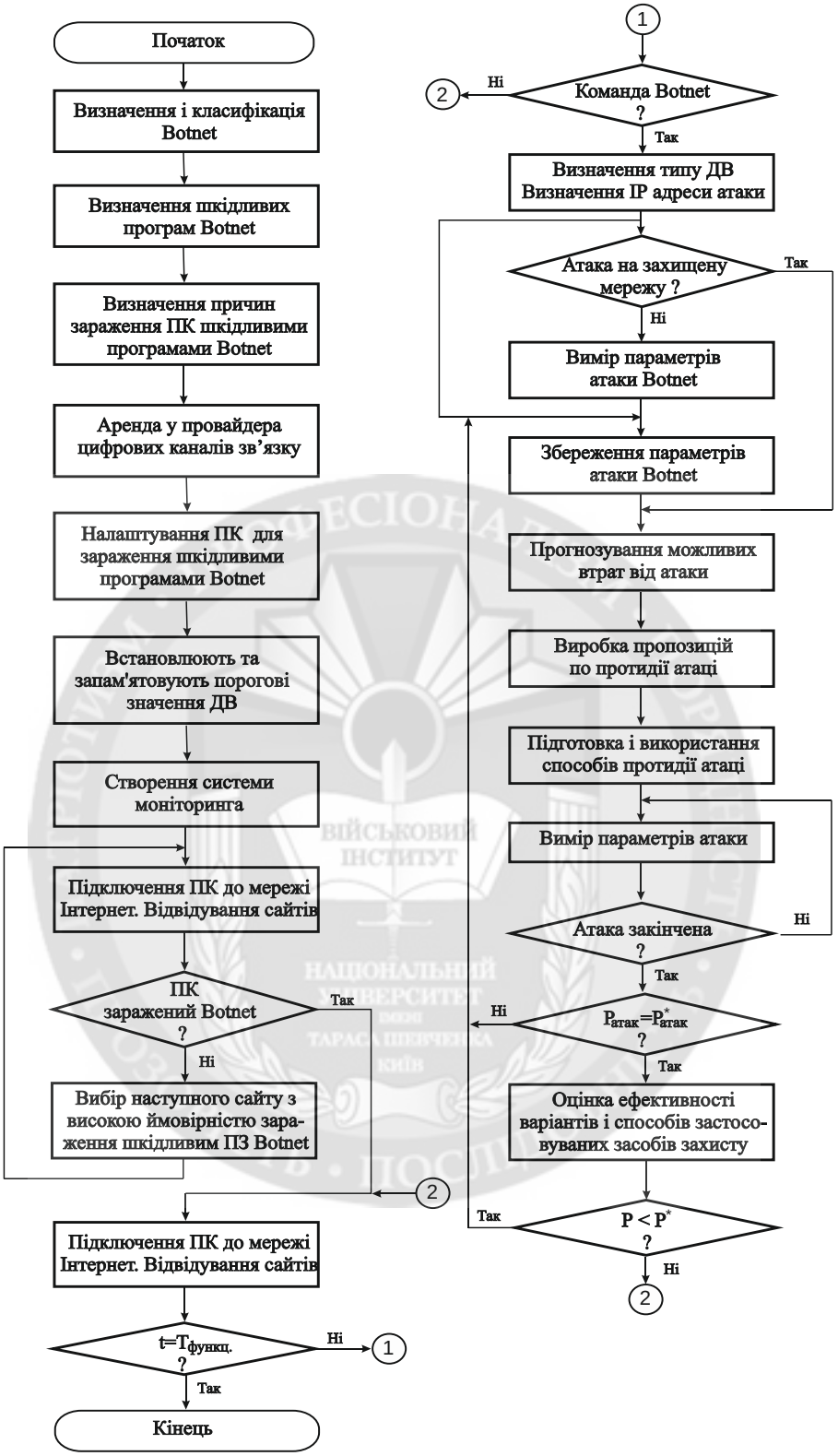

Рисунок 3 - Алгоритм раннього виявлення деструктивних впливів Botnet на мережу передачі даних

Результат 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).

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, ір-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 побудова моделей даних, обробка та пошук ознакового простору, що дозволяють алгоритмам машинного навчання з високою точністю проводити класифікаціїю даних, об'єкти та дії різних класів.

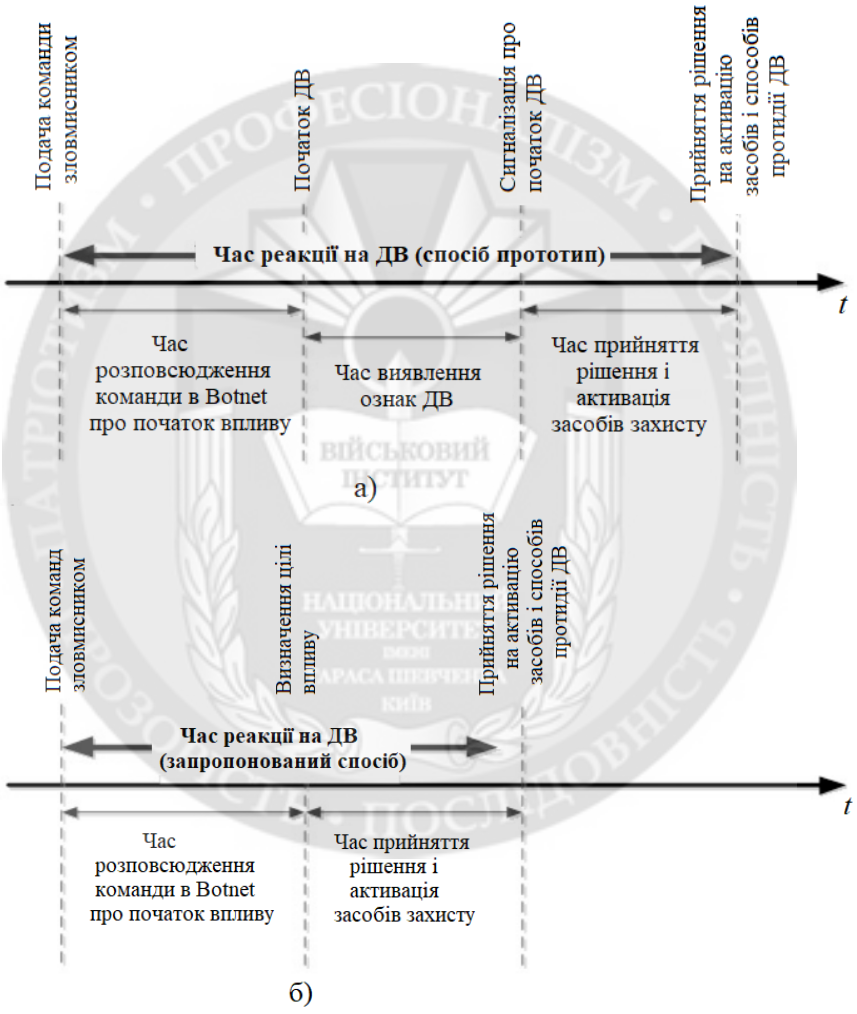

Рисунок 4 — Вирішення проблеми раннього виявлення деструктивного впливу Botnet (а) – спосіб прототип; б) -запропонований спосіб)

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. 5. Зашифрований потік даних може передаватися з робочих станцій співробітників, різних інформаційних систем, мережевих сховищ.

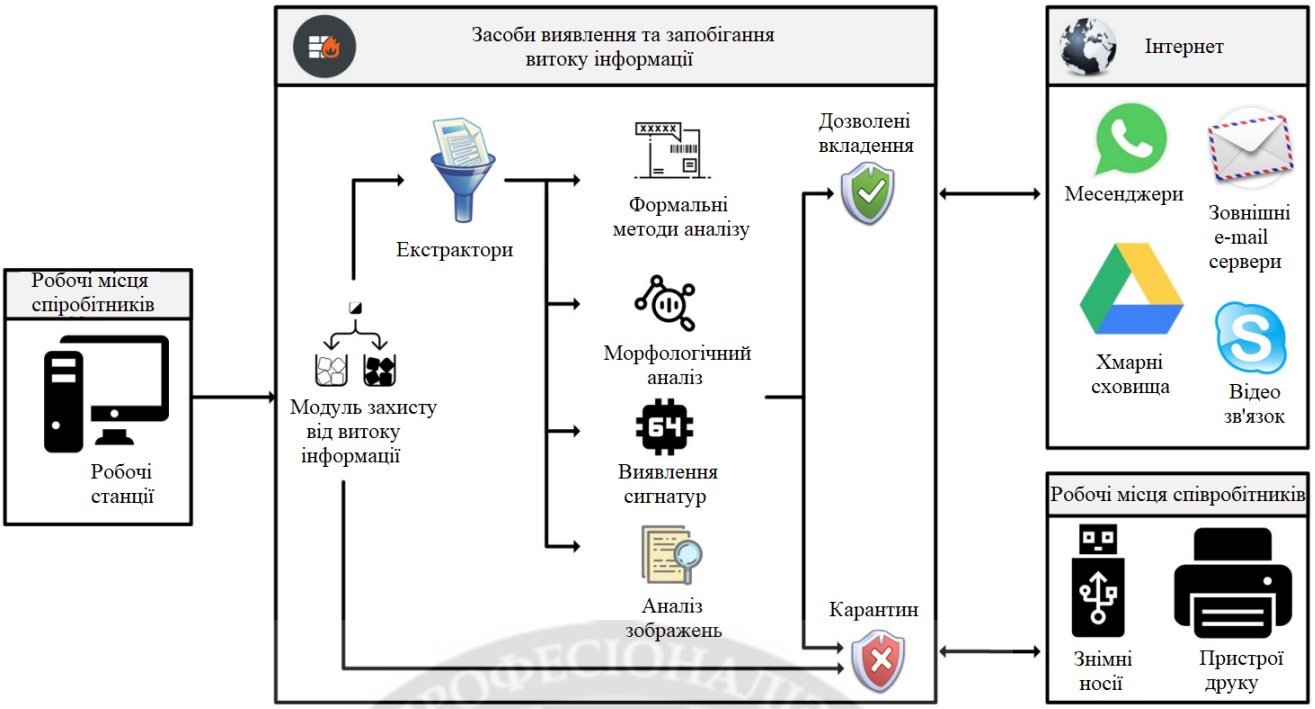

Рисунок 5 - Схема впровадження запропонованого модуля статистичного аналізу даних у DLP-системи

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. 6.

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

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. 7.

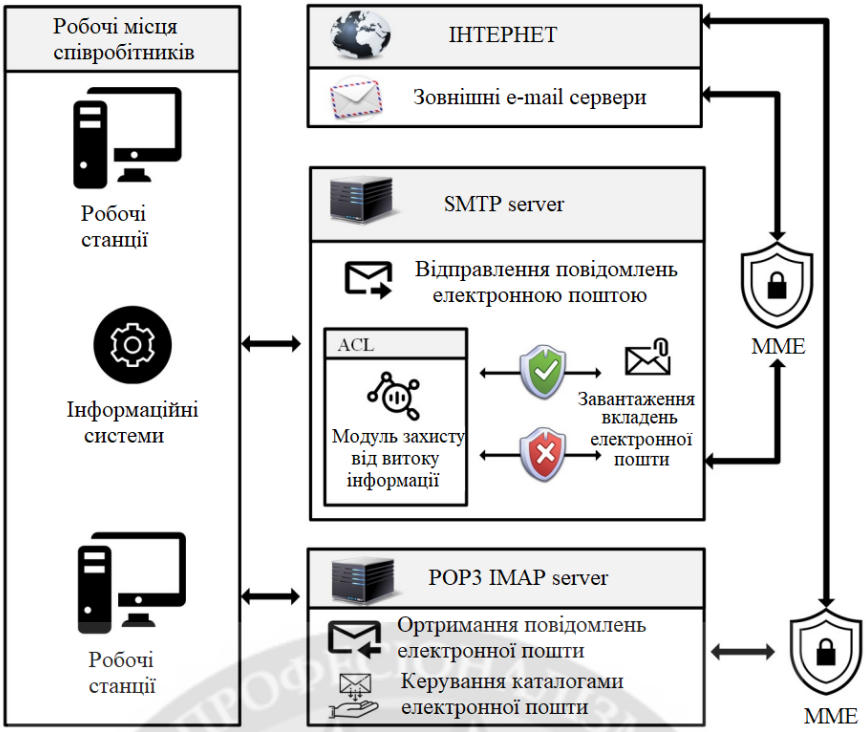

Рисунок 6 - Схема впровадження модуля статистичного аналізу даних на сервер електронної пошти методи стиснення даних

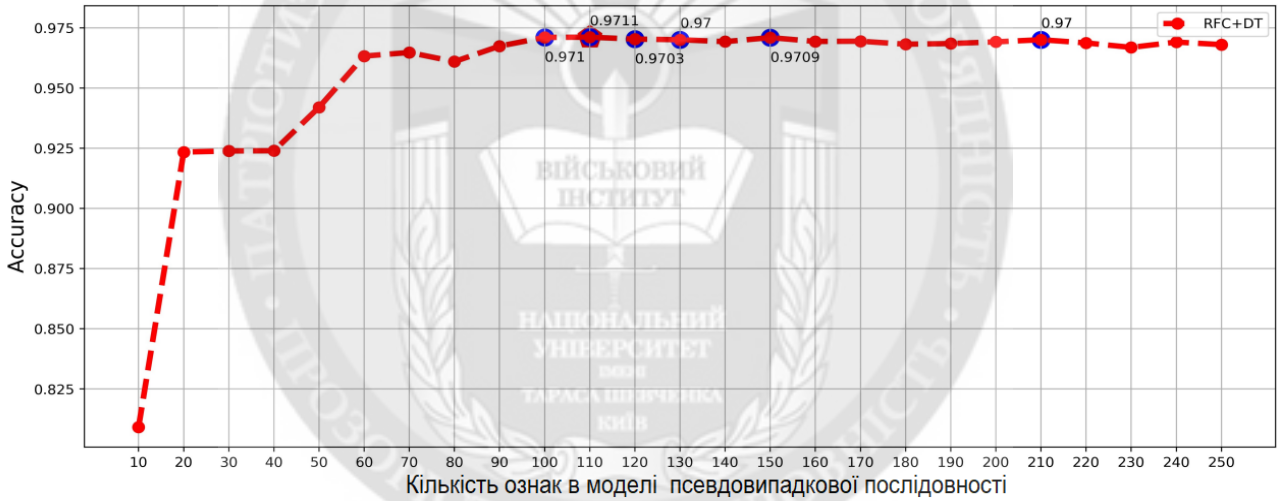

Рисунок 7 — Оцінка точності класифікатора від числа ознак моделі псевдовипадкової послідовності

Алгоритм побудови випадкового лісу відноситься до ансамблевих методів, для 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. 8 та 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. 10.

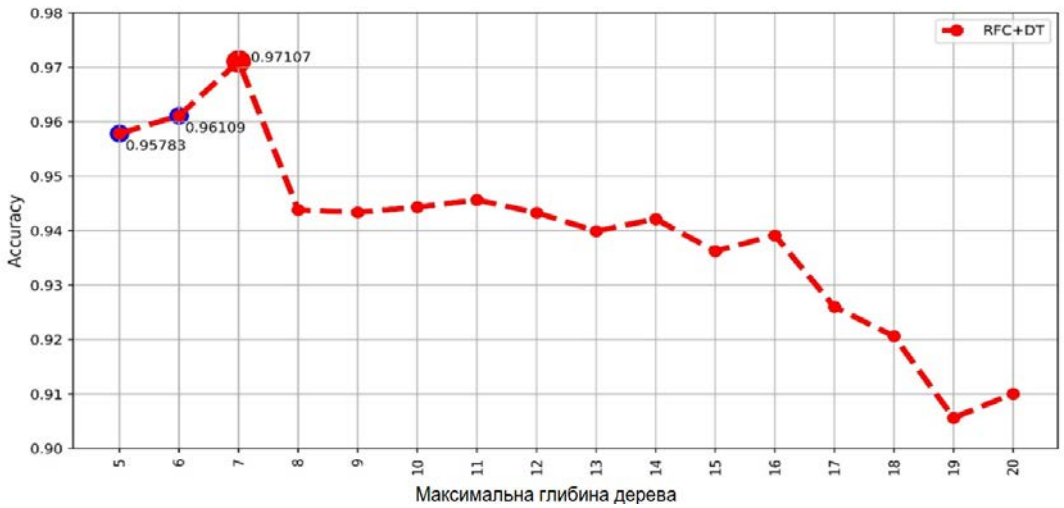

Рисунок 8 - Оцінка точності класифікатора від максимальної глибини дерев

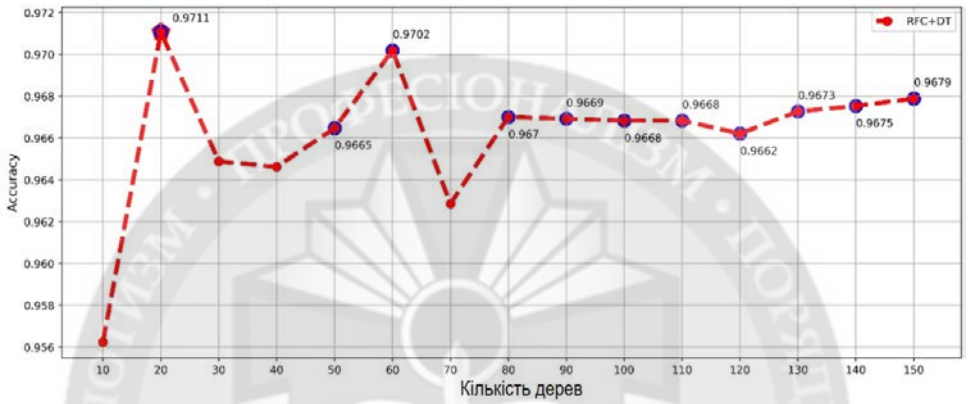

Рисунок 9 — Оцінка точності класифікатора від кількості дерев

Рисунок 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.

Висновки. Запропонований метод класифікації псевдовипадкових послідовностей, сформованих алгоритмами стиснення і шифрування інформації, враховує дискримінуючу здатність статистичних ознак даних і його реалізацію. 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 раціональне,

що дозволяє досягти класифікації псевдовипадкових послідовностей високої точності та мінімального часу виконання процедури класифікації. Обґрунтовано вибір скануючого оптимального вікна класифікатора розміром в 500 кбайт. Залежно від вимог до точності та швидкості аналізу даних запропоновано два режими роботи: сканування випадково вибраного фрагмента файла розміром 500 кб; сканування всього файлу скануючим вікном розміром 500 кб.

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.

ЛІТЕРАТУРА:

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

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

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

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

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

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

7. Ленков С.В. Функціональна модель класифікації псевдовипадкових послідовностей зашифрованих та стислих даних запобіганню витоку конфіденційної інформації / С.В. Лєнков, В.М. Джулій, І.В. Муляр, І.В. Пампуха // Збірник наукових праць Військового інституту Київського національного університету імені Тараса Шевченка. – К.: ВІКНУ, 2024. – Вип. №80. – С. 85-97.

8. 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.

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

10. 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.

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

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

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

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

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

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

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

18. 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

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

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

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

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

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

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

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

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

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

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

REFERENCES:

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

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

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

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

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

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

7. Lenkov, S.V.(2024), Funktsionalna model klasyfikatsii psevdovypadkovykh poslidovnostei zashyfrovanykh ta styslykh danykh zapobihanniu vytoku konfidentsiinoi informatsii / S.V. Lienkov, V.M. Dzhulii, I.V. Muliar, I.V. Pampukha // Zbirnyk naukovykh prats Viiskovoho instytutu Kyivskoho natsionalnoho universytetu imeni Tarasa Shevchenka. – K.: VIKNU, 2024. – Vyp. №80. – C. 85-97.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Dr. Sci. Prof. Lienkov S.V., Ph.D. Dzhuliy V.M., Ph.D.Muliar I.V.

METHOD OF CLASSIFICATION OF PSEUDO-RANDOM SEQUENCES OF COMPRESSED AND ENCRYPTED DATA TO PREVENT INFORMATION LEAKAGE

The considered task of developing a method for classifying pseudo-random sequences of protection against the leakage of confidential information based on the division of compressed and encrypted data can be used to detect network attacks on data transmission networks, in means of prevention and detection of information leakage, as well as in software products that implement services of electronic mail.

It is shown that data security threats are characterized by a set of qualitative and quantitative vector indicators, and their formalization requires the application of fuzzy set theory and discrete mathematics. It is shown that it is impossible to use expert traditional assessment methods to determine most of the considered indicators. To minimize the risk of leakage of confidential information, it is suggested to form groups of employees and calculate the risk of leakage of confidential data for each of them.

Modern means of preventing and detecting information leaks use various methods of data flow analysis. The main ones include contextual and content methods. The above methods are not able to detect

a data leak in compressed and encrypted form, and the addition of digital signatures allows you to mask encrypted data as compressed in a simple way, in the field of information security, behavioral methods of data flow analysis and machine learning algorithms have found wide use. One of the main difficulties in this situation is the construction of data models, processing and search of the feature space.

The proposed method of classifying pseudo-random sequences takes into account the discriminating ability of statistical features, it can be implemented into existing means of preventing and detecting information leaks in order to eliminate the mentioned shortcomings. An encrypted data stream can be transmitted from employee workstations, various information systems, and network storage.

To evaluate the effectiveness of the proposed method of protection against leakage of confidential data, experiments were conducted to determine the accuracy of binary classification of compressed and encrypted data depending on the types of input sequences subjected to compression procedures.

In the course of practical implementation, a quantitative assessment of the classification accuracy of pseudorandom sequences was carried out depending on the parameters of the proposed classifier. The choice of the subsequence length of nine bits is justified as the most rational value, which allows to achieve classification of pseudo-random sequences with high accuracy and minimal time for the classification procedure. The choice of the optimal scanning window of the classifier with a size of 500 kb is justified. Depending on the requirements for accuracy and speed of data analysis, two modes of operation are proposed: scanning of a randomly selected fragment of a file with a size of 500 kb; scanning the entire file with a 500 KB scanning window.

A description of the places of implementation of the proposed method of classifying pseudo-random sequences into e-mail protection subsystems, network attack detection systems, means of preventing and detecting information leaks is given. A comparative evaluation of the proposed algorithm with known analogues in the subject area of research was carried out.

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