УДК 004.056
doi:10.31799/1684-8853-2024-5-57-67
EDN: NQLXNY

Научные статьи
Articles

Выявление сетевых вторжений в промышленных киберфизических системах на основе сверточных нейронных сетей

Е. С. Новикова[а], *канд. техн. наук, доцент, orcid.org/0000-0003-2923-4954, novikova@comsec.spb.ru*
Е. О. Кузнецова[б], *магистрант, orcid.org/0009-0008-2186-8630*
С. А. Голубев[а], *аспирант, orcid.org/0009-0000-4163-5326*
[а]*Санкт-Петербургский Федеральный исследовательский центр РАН, 14-я линия В. О., 39, Санкт-Петербург, 199178, РФ*
[б]*Санкт-Петербургский государственный электротехнический университет «ЛЭТИ», Профессора Попова ул., 5, Санкт-Петербург, 197022, РФ*

Введение: *одной из наиболее сложных проблем в области обнаружения вторжений является детектирование новых, ранее неизвестных атак. В последнее время для решения этой задачи активно исследуются и применяются методики на основе глубокого обучения, поскольку они способны эффективно извлекать пространственные и временные закономерности в данных.* ***Цель:*** *разработать методику выявления сетевых атак на основе сверточных нейронных сетей для повышения сетевой безопасности промышленных киберфизических систем.* ***Результаты:*** *исследованы и систематизированы подходы к выявлению сетевых атак, основанные на представлении сетевых данных в виде двумерной матрицы анализируемых атрибутов, т. е. в виде изображения. Предложена методика выявления сетевых атак на основе сверточной нейронной сети, отличительной особенностью которой является преобразование «сырых» сетевых потоков в двумерную матрицу с последующим формированием дополнительных атрибутов, представленных текстурными признаками Харалика. Разработана архитектура нейронной сети, выполняющей анализ матричного представления сетевого трафика и вектора признаков Харалика. Для демонстрации эффективности разработанной методики выполнена серия экспериментов с использованием набора данных SWaT, описывающего функционирование системы водоочистных сооружений. В ходе экспериментов исследовалось влияние каждого компонента методики на точность обнаружения сетевых атак. Кроме того, выполнен сравнительный анализ ее эффективности с эффективностью методики обнаружения вторжений, использующей алгоритм Random Forest и описательные статистики сетевых потоков в качестве анализируемых атрибутов. Полученные результаты показали, что предложенная методика имеет высокую точность обнаружения сетевых атак, связанных с извлечением (data exfiltration) и (или) подменой передаваемых данных (data manipulation), в частности, точность повысилась на 25 % по сравнению с методикой на основе Random Forest и составила 86,3 % на исследуемом наборе SWaT.* ***Практическая значимость:*** *разработанная методика может быть использована для выявления атак, связанных с подменой передаваемых данных и (или) их извлечением.*

Ключевые слова *— промышленные киберфизические системы, выявление сетевых атак, сетевые потоки, двумерные матрицы, изображения, признаки Харалика, сверточные нейронные сети.*

Для цитирования: Новикова Е. С., Кузнецова Е. О., Голубев С. А. Выявление сетевых вторжений в промышленных киберфизических системах на основе сверточных нейронных сетей. *Информационно-управляющие системы*, 2024, № 5, с. 57–67. doi:10.31799/1684-8853-2024-5-57-67, EDN: NQLXNY

For citation: Novikova E. S., Kuznetsova E. O., Golubev S. A. Network intrusion detection based on convolutional neural networks in industrial cyber-physical systems. *Informatsionno-upravliaiushchie sistemy* [Information and Control Systems], 2024, no. 5, pp. 57–67 (In Russian). doi:10.31799/1684-8853-2024-5-57-67, EDN: NQLXNY

Введение

В настоящее время киберфизические системы получили широкое применение в различных отраслях народного хозяйства: в транспорте, электроэнергетике, промышленности, медицине и т. д. Вместе с ростом уровня информатизации этих отраслей увеличивается и число различных информационных угроз. Их реализация может вызвать как нарушение основных функций системы, так и серьезные экономические и экологические последствия. Например, в 2019 г. кибератака на норвежскую компанию Norsk Hydro ASA коснулась 35 000 сотрудников компании и привела к частичному переходу на ручное управление производством на период восстановления после атаки (https://www.hydro.com/en/global/about-hydro/company-history/2018-present/2019-cyber-attack-on-hydro/). В 2021 г. результатом кибератаки на водоочистные сооружения г. Олдсмар в США стало повышение уровня гидроксида натрия в воде (https://pcsoweb.com/21-015-detectives-investigate-computer-software-intrusion-at-oldsmar%E2%80%99s-water-treatment-plant). Кибератака в феврале 2024-го на немецкую компанию Varta по производству элементов питания привела к остановке пяти заводов, расположенных в разных частях

мира (https://www.varta-ag.com/en/about-varta/news-press/details/varta-makes-good-progress-in-solving-the-cyberattack). Таким образом, своевременное обнаружение вторжений в киберфизических системах является важной задачей. В последнее время для ее решения активно применяются модели глубокого машинного обучения, поскольку они отличаются способностью извлекать пространственные и временные закономерности в данных, что делает их привлекательными для разработки методик обнаружения новых, ранее неизвестных атак.

В статье предлагается методика обнаружения вторжений на основе сверточной нейронной сети, отличительной особенностью которой является преобразование «сырых» сетевых потоков в двумерную матрицу с последующим вычислением дополнительных атрибутов — текстурных признаков Харалика. Под сырым сетевым потоком в работе понимается сетевой поток на уровне протоколов TCP/IP, представленный в виде двоичного массива (дампов соответствующих сетевых пакетов). Такое решение позволяет выявлять скрытые зависимости между передаваемыми данными в потоке независимо от используемого сетевого протокола, поскольку не требует применения процедур по конструированию специальных признаков, для вычисления которых необходимы специальные экспертные знания. Примером таких признаков могут служить число передаваемых пакетов в потоке, число различных состояний сетевых соединений в потоке, типы используемых сетевых протоколов и т. д. Оценка эффективности разработанной методики выполнена с использованием набора данных SWaT, описывающего функционирование системы водоочистных сооружений [1].

Анализ релевантных работ

В настоящее время для обнаружения вторжений предложено большое число разнообразных подходов, в основе которых лежат методы статистического анализа данных, фрактального анализа временных рядов [2], машинного и глубокого обучения [2–6]. Например, в [4] представлена модель обнаружения вторжений в потоке данных от датчиков, которая представляет собой комбинацию сверточной нейронной сети и модулей долгой краткосрочной памяти. Х. Ван и В. Ли [5] разработали гибридную нейронную сеть DDosTC, объединяющую механизмы самовнимания и сверточные слои для обнаружения DDoS-атак в программно определяемых сетях. Андрезини и др. [6] представили многоступенчатую систему обнаружения сетевых вторжений, в которой на первом этапе применяется сверточная нейронная сеть, обученная в режиме с учителем, а на втором — два автокодировщика, обученные в режиме без учителя. Два автокодировщика используются для реконструкции нормальных и аномальных потоков, а также для формирования расширенной обучающей выборки. В [7] рассмотрен подход GraphDDoS для обнаружения низкоскоростных и высокоскоростных DDoS-атак, в его основе лежит графовая нейронная сеть, анализирующая сетевые подключения и связи между ними.

Чаще всего в таких подходах в качестве входных данных используется одномерный вектор, элементы которого представлены статистическими параметрами сетевого потока. Однако в последнее время активно развиваются методы обнаружения вторжений, в которых одномерный входной вектор данных преобразуется в двумерную матрицу, что позволяет применять двумерные свертки, которые эффективнее извлекают пространственные зависимости между атрибутами. В [8] представлена таксономия методов обнаружения вторжений, использующих преобразования сетевых данных в двумерную матрицу. В научной литературе такие подходы часто обозначаются как методы на основе анализа изображений (image-based methods), поскольку получаемая матрица может быть интерпретирована как цветное изображение или изображение в серых оттенках.

В большинстве работ [9–14] двумерная матрица строится на основе одномерного вектора статистик, вычисленных для сетевого потока. Например, в [12] векторы с сетевыми статистиками объединяются в группы по 24 объекта, для которых строится цветное изображение, полученная матрица подается на вход нейронной сети-трансформеру ViT, архитектура которой была предложена специально для анализа изображений [13]. Очевидно, что ключевым моментом в данном подходе является предположение, что сгруппированные сетевые потоки, формирующие изображение, упорядочены во времени. Другим существенным недостатком является высокая требовательность модели к вычислительным ресурсам.

В [14] предложен оригинальный способ построения цветного изображения, он состоит из трех последовательных этапов. Сначала выбираются три различных нелинейных метода снижения размерности, выполняющих проекцию многомерных данных в двумерное пространство, например t-SNE, метод главных компонент с косинусной ядерной функцией и UMAP. Далее, для каждой полученной проекции исходной обучающей выборки определяется минимальный ограничивающий прямоугольник, который затем вращается для получения финаль-

ных координат проекций каждой точки данных. Полученные проекции приводятся к единому размеру 120 × 120 и объединяются в цветное изображение, в котором каждый цветовой канал — красный, зеленый и синий — задается одной из трех проекций. Полученное изображение анализируется нейронной сетью, состоящей из двух четырехслойных сверточных сетей, объединяемых через слой конкатенации, классификация изображений осуществляется с помощью полносвязной нейронной сети.

Другой подход к преобразованию сетевых данных в изображение представлен методами, в которых двумерная матрица строится непосредственно на основе исходных сетевых данных, причем изображения могут быть построены как для пакетов [15–17], так и для потоков [18, 19]. Например, в [17] цветное изображение строится для последовательности из k пакетов. Для каждого пакета вычисляется множество вспомогательных признаков, таких как IP-адреса, порты получателя и отправителя, тип сетевого протокола, и формируется вектор, состоящий из вспомогательных признаков и первых 1458 байт пакета. Данный вектор определяет один ряд пикселей в изображении. Цвет пикселя задается направлением пакета: значения байт входящих пакетов кодируются зеленым цветом, а исходящих — красным. Значение параметра k определяется опытным путем, например, для тестового набора данных CIC-IDS2017 оптимальное количество пакетов составило девять. В качестве модели выявления сетевых аномалий используется четырехслойная сверточная нейронная сеть. Достоинством данного подхода является возможность выявлять аномальную сетевую активность в режиме, близком к реальному времени.

Показано [8], что преобразование сетевого трафика на уровне пакетов является ресурсоемкой задачей, в том числе и на этапе формирования обучающей выборки, и предложено генерировать изображения на уровне сетевых потоков. Также были исследованы различные способы компоновки пикселей в изображение, в частности с помощью кривых, заполняющих пространство. Выполненные авторами эксперименты показали, что точность сверточной нейронной сети, используемой в качестве детектора сетевых аномалий, практически не зависит от способа формирования изображения, использование прямого последовательного преобразования байт в пиксель вычислительно эффективнее, что делает его применение на практике более предпочтительным.

Следует также отметить, что предложены гибридные подходы, объединяющие представления сетевых данных на уровне пакетов и потоков [19]. В их основе лежит предположение, что анализ данных на уровне потоков позволяет выявить пространственные закономерности в сетевом трафике, а анализ данных на уровне пакетов — временные закономерности.

Таким образом, в настоящее время предложены различные методики преобразования сетевого трафика в двумерные матрицы. Их отличительной особенностью является возможность анализировать данные на уровне содержания сетевых пакетов и (или) потоков. Однако методики, выполняющие генерацию изображений на основе пакетов, вычислительно ресурсоемки в силу большого объема данных, который необходимо обработать, что делает их применение на практике нецелесообразным. Эффективность методик, в основе которых лежит преобразование сетевых потоков в изображения, исследована на наборах данных CIC-IDS2017 [12, 14, 17–19], UNSW-NB2015 [10–12, 18], в которых в основном представлены такие атаки, как отказ в обслуживании (DoS- и DDoS-атаки), сканирование портов, заражение бот-сетью. Между тем в [20] показано, что обнаружение подобных атак эффективно выполняется путем анализа статистических характеристик сетевых потоков. Следовательно, необходимо выполнить оценку эффективности методики, в основе которой лежит преобразование сетевых потоков в изображения, в задаче выявления атак другого типа, в частности атак вида извлечение данных (data exfiltration) и подмена передаваемых данных (data manipulation).

Стоит добавить, что в настоящей работе впервые предложено использовать дополнительные признаки, вычисляемые на основе сгенерированных изображений, для повышения эффективности выявления вредоносной активности в сети.

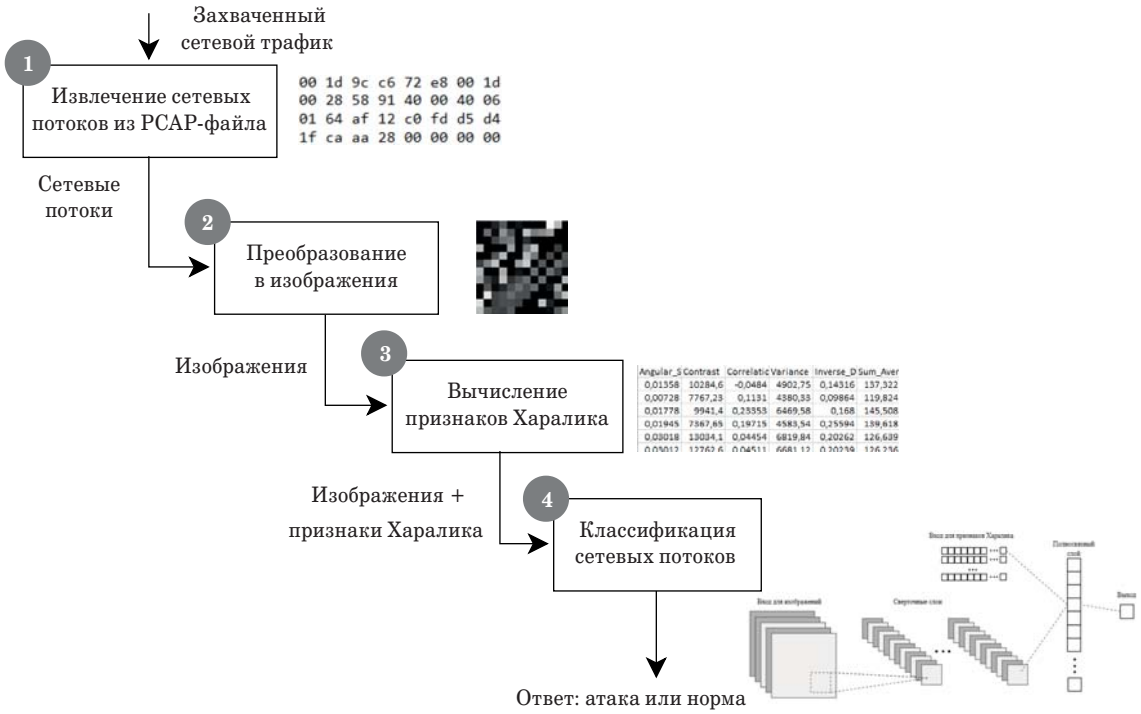
Методика обнаружения сетевых атак на основе сверточной нейронной сети

В основе разработанной методики лежат следующие предположения:

1) двумерная свертка лучше извлекает пространственные зависимости между атрибутами, чем одномерная [19];

2) непосредственный анализ содержимого сетевого потока позволяет выявлять атаки, которые не влияют на статистические характеристики потока, такие как число пакетов, средняя длина пакетов и т. д.;

3) преобразование данных в двумерную матрицу позволяет интерпретировать получаемую матрицу как изображение в оттенках серого, что дает возможность формировать новые признаки для выявления сетевых атак, которые вычисляются на основе анализа изображения.



■ ***Рис. 1.*** Схема методики к выявлению сетевых вторжений
■ ***Fig. 1.*** Scheme of the approach to network intrusion detection

Рассмотрим подробно каждый шаг представленной на рис. 1 схемы методики.

Извлечение сетевых потоков из дампа сетевого трафика

В работе используется следующее определение сетевого потока. Сетевой поток — это однонаправленная последовательность пакетов, объединенных семью общими свойствами: входным интерфейсом, IP-адресом источника, IP-адресом назначения, номером протокола IP, портом источника, портом назначения, типом IP-сервиса. При извлечении потоков также выполняется анонимизация пакетов, которая включает в себя удаление адресов. В [21] было доказано, что любые данные об адресах, такие как MAC-адрес и (или) IP-адрес источника и назначения, могут оказывать существенное влияние на эффективность обнаружения вторжений, поскольку эти признаки могут быть использованы моделью для определения класса атаки. В предлагаемом подходе анонимизация выполняется путем замены MAC- и IP-адресов на нули.

Преобразование сетевого потока в изображение

Перед определением функции преобразования сетевого потока в изображение необходимо задать три основных параметра: 1) размер изображения; 2) цветовой режим (оттенки серого или RGB); 3) способ компоновки атрибутов (пикселей) в изображении.

Анализ соответствующих исследований показал, что не существует единого подхода к определению размера изображения, хотя этот параметр оказывает решающее влияние на точность модели. Часто для визуализации данных выбирается фиксированное число байт сетевого пакета, а размер изображения определяется с учетом формата входных данных для нейронной сети [16, 21, 22]. В настоящей работе предлагается для определения размера изображения $n \times n$ использовать формулу $n = \text{ceil}(\sqrt{P_{stat}})$, где ceil — функция округления до ближайшего целого, а P_{stat} — статистический показатель, определяемый на основе статистического анализа распределения длин потоков в обучающей выборке. В качестве P_{stat} может быть выбран средний размер сетевого потока, максимальный или минимальный размер потока, медиана и т. д. В настоящей работе параметр P_{stat} задается средним значением размера потока.

В разработанной методике изображения строятся в оттенках серого. Такие изображения имеют только один канал, который передает информацию об интенсивности света и принимает значения в диапазоне от 0 до 255. В этом случае алгоритм преобразования сетевых пакетов является довольно простым: каждый сетевой поток представляется в виде двоичной последователь-

ности, которая делится на байты. Затем каждый байт преобразуется в уровень серого в соответствии со следующим правилом (рис. 2):

$0 \times 00 \rightarrow 0$ (черный), ..., $0 \times FF \rightarrow 255$ (белый).

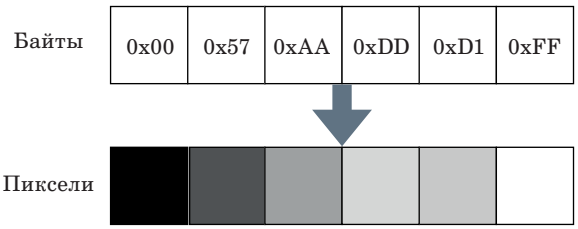
В этом случае не требуется извлекать такие атрибуты, как тип протокола, тип сервиса и др.

Существует несколько подходов к заполнению пространства изображения [8]. В нашей работе используется линейное заполнение, при котором каждый байт сетевого потока последовательно преобразуется в пиксель, начиная с верхнего левого угла изображения и заканчивая нижним правым углом. Когда строка заполняется, обрабатывается следующая строка. Это действие повторяется до тех пор, пока изображение не будет завершено. Если пакеты имеют меньшую длину, чем число элементов матрицы, то оставшиеся пиксели заполняются 0×00.

Вычисление признаков Харалика

В качестве дополнительных анализируемых параметров используются признаки Харалика [23]. Они применяются для описания текстуры изображения и позволяют количественно оценить и описать визуальные и тактильные свойства поверхностей.

Признаки Харалика вычисляются с помощью специальной матрицы совместной встречаемости на уровне серого (Gray-Level Co-Occurrence Matrix, GLCM), которая показывает, как часто пиксель со значением интенсивности (уровня серого) i встречается в определенном пространственном соотношении с пикселем со значением j. Чтобы составить матрицу GLCM, в соответствующие элементы записывается число раз, когда пиксели определенной интенсивности находились рядом друг с другом. На рис. 3 представлена схема расчета элементов матрицы GLCM: пиксель интенсивности *3* находится справа от пикселя *2* один раз, поэтому элемент матрицы [3, 2] получает значение *1*, а комбинация пикселей *3–2* встречается дважды, поэтому элемент матрицы [2, 3] выставляется в значение *2*.



■ ***Рис. 2.*** Кодирование байт в оттенки серого
■ ***Fig. 2.*** Byte encoding in grayscale color

Очевидно, что оценить комбинации соседних пикселей можно не только слева направо, но и в других направлениях: справо налево, сверху вниз и по двум диагоналям (рис. 4). Таким образом, получается четыре матрицы GLCM, которые можно использовать для расчета признаков Харалика.

Признаки Харалика являются вторичными текстурными признаками, их перечень приведен в табл. 1.

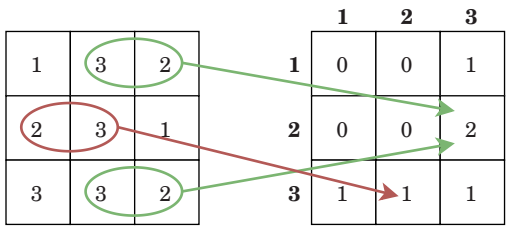
Классификация сетевых вторжений

Задача выявления сетевых вторжений нами рассматривается как задача бинарной классификации, в которой вторжения представлены потоками, соответствующими различным типам сетевых атак и обозначенными одной меткой — 1 (атака). Для ее решения была использована модель сверточной нейронной сети, которая является наиболее распространенной и эффективной архитектурой для анализа изображений.

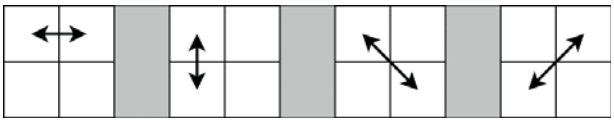
Поскольку для классификации сетевых потоков предлагается анализировать пиксели изображения и признаки Харалика, было реализовано два входных слоя для модели. Изображение обрабатывается тремя сверточными слоями, после каждого из которых следует подвыборочный слой с функцией max pool. Выход третьего подвыборочного слоя объединяется с входом для признаков Харалика, после чего подается для дальнейшей обработки на полносвязный слой (рис. 5).

Особенности программной реализации подхода

Исходные сетевые данные в формате PCAP-файла разбиваются на сетевые потоки с помощью специального программного инструмента



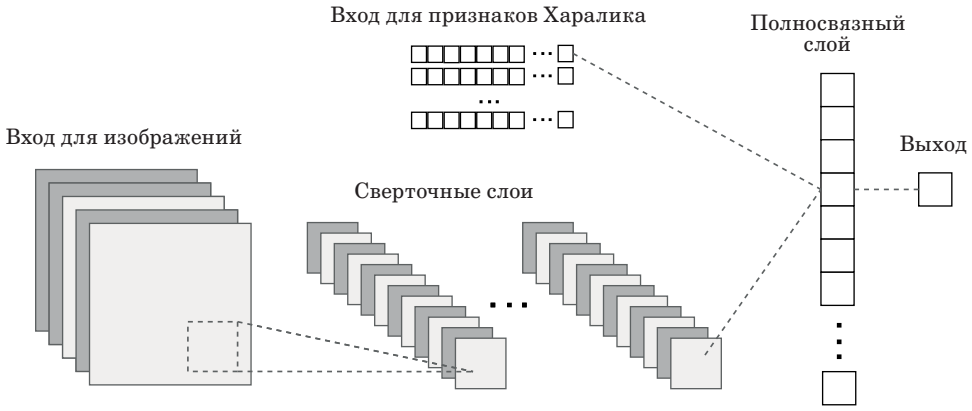
■ ***Рис. 3.*** Вычисление матрицы GLCM
■ ***Fig. 3.*** Calculation of GLCM matrix



■ ***Рис. 4.*** Способы вычисления матрицы GLCM
■ ***Fig. 4.*** Ways to calculate GLCM matrix

■ ***Таблица 1.*** Описание текстурных признаков Харалика [23]
■ ***Table 1.*** Description of the Haralick texture features [23]

Признак	Описание
Второй угловой момент (Angular Second Moment)	Измеряет локальную однородность уровней серого. Если пиксели очень похожи, значение будет большим
Контраст (Contrast)	Изменение интенсивности или уровня серого между опорным пикселем и соседним с ним. Большой контраст отражает большие различия в интенсивности в GLCM
Корреляция (Correlation)	Показывает линейную зависимость значений уровня серого в матрице GLCM
Однородность (Variance)	Измеряет неоднородность текстуры
Обратный момент разностей (Inverse Difference Moment)	Измеряет однородность текстуры
Сумма средних (Sum Average)	Измеряет сумму средних значений всех пикселей
Суммарная неоднородность (Sum Variance)	Признак неоднородности, который сильно коррелирует со статистической переменной первого порядка, такой как стандартное отклонение. Дисперсия увеличивается, когда значения уровня серого отличаются от их среднего значения
Суммарная энтропия (Sum Entropy)	Характеризует неоднородность изображения или сложность текстуры
Энтропия (Entropy)	Измеряет случайность интенсивности пикселей
Неоднородность разностей (Difference Variance)	Показывает неоднородность изображения
Энтропия разностей (Difference Entropy)	Отражает уровень случайности, отсутствие структуры или порядка в контрастности изображения
Информационные показатели корреляции 1, 2	Измеряют корреляцию параметров матрицы с использованием дополнительных методов



■ ***Рис. 5.*** Архитектура разработанной нейронной сети
■ ***Fig. 5.*** Architecture of the proposed neural network

NetFlow2Image (https://github.com/EveNovikova/FedIDSExplorer), разработанного на языке программирования Python с использованием библиотеки Scapy (https://scapy.net/). Инструмент позволяет сразу разметить дамп на аномальные сетевые потоки и норму, используя файл с разметкой в формате JSON, в котором указываются IP-адреса атакующих и атакуемых хостов и начальное и конечное время атаки. Полученные сетевые потоки дальше преобразуются в черно-белые изображения с помощью библиотек NumPy и Pillow (https://pypi.org/project/pillow/). Пользователь имеет возможность задать различные настройки формирования изображения. Полученные изображения сетевых потоков сохраняются в формате PNG, при этом формиру-

ется иерархия директорий (аномальные сетевые потоки или норма) для дальнейшего проведения обучения нейронных сетей. Данная утилита также использует библиотеку CICFlowMeter (https://github.com/ahlashkari/CICFlowMeter), которая вычисляет различные статистики для сетевого потока, наиболее часто применяемые для выявления аномальной сетевой активности.

Расчет признаков Харалика осуществляется с помощью библиотеки Mahotas (https://mahotas.readthedocs.io/en/latest/), а для разработки нейронной сети была использована программная библиотека TensorFlow (https://www.tensorflow.org/?hl=ru).

Экспериментальная оценка

Целью эксперимента являлось определение эффективности предложенного подхода к обнаружению сетевых атак, для чего был разработан следующий сценарий.

На первом этапе выполнялась оценка влияния направления расчета признаков Харалика на точность обнаружения сетевых вторжений.

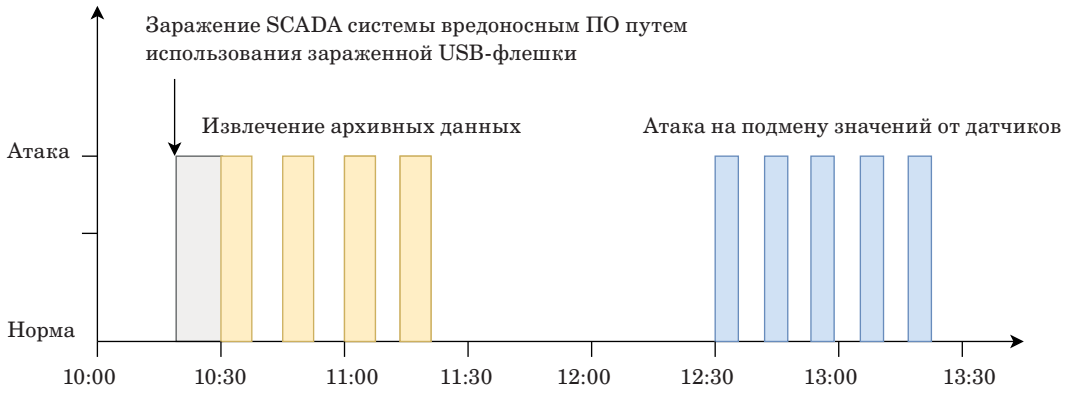
На втором этапе оценивалось влияние использования признаков Харалика в качестве дополнительных атрибутов. Для этого была проведена серия экспериментов, в которой сравнивалась точность обнаружения вторжений с использованием только сверточной нейронной сети и с использованием сверточной сети и признаков Харалика. Кроме того, для оценки целесообразности применения преобразования сетевого трафика в изображения было выполнено сравнение с моделью обнаружения вторжений на основе алгоритма Random Forest, обученного на классических признаках, описывающих статистические параметры сетевых потоков. Выбор этой модели объясняется тем, что она наиболее часто используется в задачах обнаружения сетевых атак.

Описание набора данных и типов атак

В качестве анализируемого набора данных был выбран набор SWaT версии 2019 г. [1], который получен с использованием водоочистительного полигона, созданного Центром исследований кибербезопасности iTrust Сингапурского университета технологий и дизайна в 2019 г. Этот полигон является уменьшенной копией водоочистных сооружений и моделирует современный процесс очистки воды. В состав полигона SWaT входят водоочистительное оборудование, многочисленные ПЛК, SCADA, автоматизированное рабочее место оператора и хранилище данных от технологического процесса.

Версия набора данных SWaT.A6_Dec 2019 состоит из PCAP-файлов с дампом сетевого трафика, файла в формате csv с показаниями датчиков и файла с описанием выполненных атак. PCAP-файлы содержат сетевой трафик, характеризующий как нормальный, так и аномальный режим работы полигона. Вредоносные сценарии представлены двумя типами атак: извлечением исторических данных (exfiltrate historian data) и подменой показателей сенсоров (disrupt sensor readings and process).

Атака на извлечение данных представляет собой преднамеренную несанкционированную передачу данных технологического процесса. Атака на подмену показателей датчиков относится к категории атак, целью которых является искажение или подмена данных, получаемых от оборудования. Последнее может быть выполнено путем физического вмешательства в работу датчика или через программное обеспечение, зараженное вредоносным кодом. Подобные атаки могут привести к неправильной работе систем управления и мониторинга, что особенно критично в промышленных и инфраструктурных объектах, где точность данных от физических устройств имеет первостепенное значение. Временная диаграмма атакующих воздействий,



■ ***Рис. 6.*** Схема атакующих сценариев в наборе данных SWaT.A6_Dec 2019
■ ***Fig. 6.*** Schema of attack scenarios in the SWaT.A6_Dec 2019 dataset

представленных в наборе SWaT.A6_Dec 2019, приведена на рис. 6 [1].

Параметры экспериментов

При обучении сверточной нейронной сети были использованы следующие настройки:

1) оптимизатор: Adam;
2) функция потерь: binary_crossentropy;
3) количество эпох обучения: 10.

Для обучения модели Random Forest настройка параметров осуществлялась с помощью функции GridSearchCV. В результате были использованы следующие параметры: n_estimators = 5 (число деревьев), max_depth = 9 (максимальная глубина деревьев), max_leaf_node = 9 (максимальное число листьев в дереве).

При выполнении эксперимента данные были анонимизированы: из анализа были исключены данные о IP-адресе сетевого потока. Для обеспечения баланса классов был использован механизм оверсэмплинга. Во всех экспериментах разделение на тренировочную и тестовую выборки производилось случайным образом в соотношении 80:20.

Оценка эффективности обнаружения вторжений осуществлялась с помощью метрик точность (precision), полнота (recall) и F1-мера.

Точность отражает долю объектов, которые действительно принадлежат данному классу относительно всех объектов, которым модель сопоставила этот класс. В рамках задачи обнаружения атак высокая точность означает, что система способна в большинстве случаев корректно детектировать атаки с сохранением низкого уровня ложных срабатываний. Полнота — доля выявленных моделью объектов, принадлежащих классу, относительно всех объектов этого класса. Высокое значение метрики полноты означает, что система способна в большинстве случаев корректно детектировать атаки с сохранением низкого уровня пропуска атак. F1-мера представляет собой среднее гармоническое между вышеуказанными метриками и выражается формулой

$$F1 = 2 \cdot (Precision \cdot Recall) / (Precision + Recall).$$

Анализ полученных результатов

Результаты точности выявления атак для разных способов формирования признаков Харалика представлены в табл. 2.

Лучшие результаты по всем метрикам были получены для угла поворота 45°. При данном способе извлечения признаков Харалика полнота обнаружения вторжений составила 98 %, а точность — 77 %.

■ ***Табл. 2.*** Результаты точности выявления атак при разных углах расчета признаков Харалика

■ ***Table 2.*** Experimental results for different angles of Haralick features

Угол расчета признаков Харалика	Precision	Recall	F1-мера
0	0,67	0,93	0,78
45	**0,77**	**0,98**	**0,86**
90	0,72	0,92	0,81
135	0,74	0,91	0,82
Усредненный	0,68	0,78	0,73

■ ***Таблица 3.*** Результаты экспериментов для различных моделей обнаружения вторжений

■ ***Table 3.*** Experimental results for different intrusion detection models

Тип модели обнаружения вторжений	Precision	Recall	F1-мера
Сверточная нейронная сеть (только изображения)	0,69	0,93	0,79
Сверточная нейронная сеть (только изображения + признаки Харалика)	**0,77**	**0,98**	**0,86**
Random Forest (статистики)	0,65	0,77	0,61

Результаты второго этапа эксперимента представлены в табл. 3.

Из нее следует, что сверточная нейронная сеть, обученная на сырых данных сетевых потоков, дает лучшие результаты обнаружения таких атак, как извлечение данных и модификация данных сенсора. Можно предположить, что это связано с тем, что такие атаки затрагивают в первую очередь содержание передаваемых данных и в меньшей степени влияют на статистические параметры сетевых потоков, как это происходит в случае атак типа сканирование портов и отказ в обслуживании.

Исследование влияния признаков Харалика на точность обнаружения сетевых атак показало, что их использование дает повышение точности решения задачи на 7 % по сравнению со сверточной нейронной сетью, обученной только на изображениях, и почти на 25 % по сравнению с моделью Random Forest. Следует отметить, что их применение позволяет в большей мере повысить точность, т. е. снизить число ложноположительных срабатываний, что важно при практическом применении разработанных моделей.

Однако вычисление признаков Харалика является достаточно ресурсоемкой задачей, вычислительная сложность которой прямо пропорциональна размеру анализируемого изображения. Эта проблема дает дальнейшее направление исследованиям, связанным как с определением временных показателей рассмотренного подхода, так и с исследованием других текстурных признаков, обладающих более высокой вычислительной эффективностью.

Заключение

В последнее время для обнаружения атак и аномалий в киберфизических системах было предложено большое число методов, в основе которых лежат модели машинного обучения, включая глубокие нейронные сети. В настоящей статье предложена методика выявления сетевых атак, отличающаяся способом преобразования сетевых потоков в двумерную матрицу с последующим формированием признаков Харалика. Детально представлены ее основные шаги: генерация изображения в оттенках серого на основе сетевого потока, вычисление признаков Харалика, классификация объекта с помощью сверточной нейронной сети.

Хотя некоторые исследователи считают, что достаточно использовать статистические параметры сетевых потоков [20, 22], а в преобразовании исходных сетевых данных в изображения нет необходимости, проведенные эксперименты с данными от тестового полигона водоочистных сооружений показали, что этот способ подготовки входных данных позволяет эффективно обнаруживать такие атаки, как подмена передаваемых данных и (или) их извлечение даже при использовании простых сверточных нейронных сетей. Применение дополнительных признаков, которые оценивают текстуру формируемых изображений, позволяет снизить число ложноположительных срабатываний и тем самым повысить точность обнаружения сетевых атак, в частности на промышленные киберфизические системы. Кроме того, благодаря тому, что исследуемые входные данные формируются путем преобразования бинарного вектора в числовую матрицу, предлагаемая методика может считаться независимой от используемого сетевого протокола и применяться для анализа сетевого трафика, передаваемого по любому сетевому протоколу, основанному на TCP/IP, например по промышленному протоколу Modbus TCP.

Дальнейшее направление исследований связано с оценкой вычислительной эффективности разработанного подхода, апробацией на других наборах данных, сформированных для других систем и других сетевых протоколов, и анализом других текстурных признаков в качестве дополнительных анализируемых атрибутов. Также в задачи будущих исследований включен поиск и анализ других архитектур нейронных сетей, в частности одномерных сверточных сетей.

Финансовая поддержка

Работа выполнена при поддержке гранта Российского научного фонда № 23-11-20024 (https://rscf.ru/project/23-11-20024/) и Санкт-Петербургского научного фонда в СПб ФИЦ РАН.

Литература

1. **Goh J., Adepu S., Junejo K. N., Mathur A.** A dataset to support research in the design of secure water treatment systems. *Critical Information Infrastructures Security*, 2017, vol. 10242, pp. 88–99. doi:10.1007/978-3-319-71368-7_8
2. **Kotenko I., Saenko I., Kribel A., Lauta A.** A technique for early detection of cyberattacks using the traffic self-similarity property and a statistical approach. *Proc. of 29th Euromicro Intern. Conf. on Parallel, Distributed and Network-Based Processing, PDP 2021*, Virtual, Valladolid, 10–12 March 2021, pp. 281–284. doi:10.1109/PDP52278.2021.00052
3. **Branitskiy A., Kotenko I., Saenko I.** Applying machine learning and parallel data processing for attack detection in IoT. *IEEE Transactions on Emerging Topics in Computing*, 2021, vol. 9, pp. 1642–1653. doi:10.1109/TETC.2020.3006351
4. **Gaber T., Awotunde J., Torky M., Ajagbe S., Hammoudeh M., Li W.** Metaverse-IDS: Deep learning-based intrusion detection system for Metaverse-IoT networks. *Internet of Things*, 2023, vol. 24, no. 100977. doi:10.1016/j.iot.2023.100977
5. **Wang H., Li W.** DDosTC: A transformer-based network attack detection hybrid mechanism in SDN. *Sensors*, 2021, vol. 21, iss. 15, pp. 5047. doi:10.3390/s21155047
6. **Andresini G., Appice A., Mauro N. D., Loglisci C., Malerba D.** Multi-channel deep feature learning for intrusion detection. *IEEE Access*, 2020, vol. 8, pp. 53346–53359. doi:10.1109/ACCESS.2020.2980937
7. **Li Y., Li R., Zhou Z., Guo J., Yang W., Du M., Liu Q.** GraphDDoS: Effective DDoS attack detection using graph neural networks. 2022 *IEEE 25th Intern. Conf. on Computer Supported Cooperative Work in Design, CSCWD*, 2022, Hangzhou, China, pp. 1275–1280. doi:10.1109/CSCWD54268.2022.9776097

8. **Golubev S., Novikova E.** Transformation of network flow data into images for intrusion detection using convolutional neural networks. *2023 Intern. Russian Automation Conf. (RusAutoCon)*, Sochi, Russian Federation, 2023, pp. 948–952. doi:10.1109/RusAutoCon58002.2023.10272890
9. **Wang Q., Zhao W., Ren J.** Intrusion detection algorithm based on image enhanced convolutional neural network. *J. Intell. Fuzzy Syst.*, 2021, vol. 41, no. 1, pp. 2183–2194. doi:10.3233/JIFS-210863
10. **Masum M., Shahriar H., Haddad H. M.** A transfer learning with deep neural network approach for network intrusion detection. *International Journal of Intelligent Computing Research (IJICR)*, 2021, vol. 12, pp. 087–1095. doi:10.20533/ijicr.2042.4655.2021.0132
11. **Noever D. A., Noever S. E. M.** Image classifiers for network intrusions. *CoRR*, 2021, arXiv:2103.07765. https://arxiv.org/abs/2103.07765 (дата обращения: 05.04.2024).
12. **Ho C. M. K., Yow K.-C., Zhu Z., Aravamuthan S.** Network intrusion detection via flow-to-image conversion and vision transformer classification. *IEEE Access*, 2022, vol. 10, pp. 97780–97793. doi:10.1109/ACCESS.2022.3200034
13. **Dosovitskiy A., Beyer L., Kolesnikov A., Weissenborn D., Zhai X., Unterthiner T., Dehghani M., Minderer M., Heigold G., Gelly S., Uszkoreit J., Houlsby N.** An image is worth 16x16 words: Transformers for image recognition at scale. 2020, arXiv:2010.11929. https://doi.org/10.48550/arXiv.2010.11929 (дата обращения: 05.04.2024).
14. **Kim T., Pak W.** Deep learning-based network intrusion detection using multiple image transformers. *Appl. Sci.*, 2023, vol. 13, no. 2754. doi:10.3390/app13052754
15. **Hosler R., Sundar A., Zou X., Li F., Gao T.** Unsupervised deep learning for an image based network intrusion detection system. *Proc. of 2023 IEEE Global Communications Conf.*, Kuala Lumpur, Malaysia, 2023, pp. 6825–6831. doi: 10.1109/GLOBECOM54140.2023.10437636
16. **Golubev S., Novikova E., Fedorchenko E.** Image-based approach to intrusion detection in cyber-physical objects. *Information*, 2022, vol. 13, iss. 12, pp. 553. doi:10.3390/info13120553
17. **Ghadermazi J., Shah A., Bastian N.** Towards real-time network intrusion detection with image-based sequential packets representation. *IEEE Transactions on Big Data*, 2024, no. 01, pp. 1–17. doi:10.1109/TBDATA.2024.3403394
18. **Zhang X., Chen J., Zhou Y., Han L., Lin J.** A multiple-layer representation learning model for network-based attack detection. *IEEE Access*, 2019, vol. 7, pp. 91992–92008. doi:10.1109/ACCESS.2019.2927465
19. **Yu L., Dong J., Chen L., Li M., Xu B., Li Z., Qiao L., Liu L., Zhao B., Zhang B.** Pbcnn: Packet bytes-based convolutional neural network for network intrusion detection. *Computer Networks*, 2021, vol. 194, no. 108117. doi:10.1016/j.comnet.2021.108117
20. **Sharafaldin I., Lashkari A. H., Ghorbani A. A.** Toward generating a new intrusion detection dataset and intrusion traffic characterization. *Proc of Intern. Conf. on Information Systems Security and Privacy*, Funchal, Madeira — Portugal, January 22–24, 2018, pp. 108–116.
21. **Sun Y., Esaki H., Ochiai H.** Adaptive intrusion detection in the networking of large-scale LANs with segmented federated learning. *IEEE Open Journal of the Communications Society*, 2021, vol. 2, pp. 102–112. doi:10.1109/OJCOMS.2020.3044323
22. **Wu P., Guo H., Buckland R.** A transfer learning approach for network intrusion detection. *Proc. of 2019 IEEE 4th Intern. Conf. on Big Data Analytics (ICBDA)*, Suzhou, China, 2019, pp. 281–285. doi:10.1109/ICBDA.2019.8713213
23. **Haralick R., Shanmugam K., Dinstein I.** Textural features for image classification. *IEEE TSMC*, 1973, vol. 3, iss. 6, pp. 610–621.

UDC 004.056
doi:10.31799/1684-8853-2024-5-57-67
EDN: NQLXNY

Network intrusion detection based on convolutional neural networks in industrial cyber-physical systems

E. S. Novikova[a], PhD, Tech., Associate Professor, orcid.org/0000-0003-2923-4954, novikova@comsec.spb.ru
E. O. Kuznetsova[b], Master Student, orcid.org/0009-0008-2186-8630
S. A. Golubev[a], Post-Graduate Student, orcid.org/0009-0000-4163-5326
[a]St. Petersburg Federal Research Center of the RAS, 39, 14th Line, 199178, Saint-Petersburg, Russian Federation
[b]Saint-Petersburg Electrotechnical University «LETI», 5, Prof. Popov St., 197376, Saint-Petersburg, Russian Federation

Introduction: One of the most challenging problems in intrusion detection is the detection of new, previously unknown attacks. Recently, deep learning techniques have been extensively researched and applied to this problem because of their ability to efficiently extract spatial and temporal patterns in data. **Purpose:** To develop a methodology for detecting network attacks based on convolutional neural networks to improve network security of industrial cyber-physical systems. **Results:** We investigate and systematize approaches to detecting network attacks based on the representation of network data in the form of a two-dimensional matrix of analyzed attributes, i.e. in the form of an image. We propose a new approach to the detection of network attacks based on convolutional neural network, the distinctive feature of this is the transformation of "raw" network flows into a two-dimensional matrix with the subsequent formation of

additional attributes represented by Haralick texture features. We develop the architecture of a neural network that analyzes the matrix representation of network traffic and Haralick feature vector. To demonstrate the effectiveness of the developed approach, we perform a series of experiments using the SWaT dataset describing the operation of a water treatment plant system. During the experiments, we have investigated the impact of each component of the approach on the detection accuracy of network attacks. In addition, we perform a comparative performance analysis with an intrusion detection method using Random Forest algorithm and descriptive statistics of network flows as analyzed attributes. The results show that the proposed technique has a high accuracy in detecting network attacks related to data exfiltration and/or data manipulation, in particular, it has improved by 25% as compared to the Random Forest-based method and equals 86.3% on the SWaT set. **Practical relevance:** The developed methodology can be used to detect attacks related to spoofing of transmitted data and/or their extraction.

Keywords — industrial cyber-physical systems, network attack detection, network flows, two-dimensional matrices, images, Haralick features, convolutional neural networks.

For citation: Novikova E. S., Kuznetsova E. O., Golubev S. A. Network intrusion detection based on convolutional neural networks in industrial cyber-physical systems. *Informatsionno-upravliaiushchie sistemy* [Information and Control Systems], 2024, no. 5, pp. 57–67 (In Russian). doi:10.31799/1684-8853-2024-5-57-67, EDN: NQLXNY

Financial support

The research is supported by the grant of Russian Science Foundation No. 23-11-20024 (https://rscf.ru/project/23-11-20024/), and St. Petersburg Science Foundation.

References

1. Goh J., Adepu S., Junejo K. N., Mathur A. A dataset to support research in the design of secure water treatment systems. *Critical Information Infrastructures Security*, 2017, vol. 10242, pp. 88–99. doi:10.1007/978-3-319-71368-7_8
2. Kotenko I., Saenko I., Kribel A., Lauta A. A technique for early detection of cyberattacks using the traffic self-similarity property and a statistical approach. *Proc. of 29th Euromicro Intern. Conf. on Parallel, Distributed and Network-Based Processing*, *PDP 2021*, Virtual, Valladolid, 10–12 March 2021, pp. 281–284. doi:10.1109/PDP52278.2021.00052
3. Branitskiy A., Kotenko I., Saenko I. Applying machine learning and parallel data processing for attack detection in IoT. *IEEE Transactions on Emerging Topics in Computing*, 2021, vol. 9, pp. 1642–1653. doi:10.1109/TETC.2020.3006351
4. Gaber T., Awotunde J., Torky M., Ajagbe S., Hammoudeh M., Li W. Metaverse-IDS: Deep learning-based intrusion detection system for Metaverse-IoT networks. *Internet of Things*, 2023, vol. 24, no. 100977. doi:10.1016/j.iot.2023.100977
5. Wang H., Li W. DDosTC: A transformer-based network attack detection hybrid mechanism in SDN. *Sensors*, 2021, vol. 21, iss. 15, pp. 5047. doi:10.3390/s21155047
6. Andresini G., Appice A., Mauro N. D., Loglisci C., Malerba D. Multi-channel deep feature learning for intrusion detection. *IEEE Access*, 2020, vol. 8, pp. 53346–53359. doi:10.1109/ACCESS.2020.2980937
7. Li Y., Li R., Zhou Z., Guo J., Yang W., Du M., Liu Q. Graph-DDoS: Effective DDoS attack detection using graph neural networks. 2022 *IEEE 25th Intern. Conf. on Computer Supported Cooperative Work in Design, CSCWD*, 2022, Hangzhou, China, pp. 1275–1280. doi:10.1109/CSCWD54268.2022.9776097
8. Golubev S., Novikova E. Transformation of network flow data into images for intrusion detection using convolutional neural networks. *2023 Intern. Russian Automation Conf. (RusAutoCon)*, Sochi, Russian Federation, 2023, pp. 948–952. doi:10.1109/RusAutoCon58002.2023.10272890
9. Wang Q., Zhao W., Ren J. Intrusion detection algorithm based on image enhanced convolutional neural network. *J. Intell. Fuzzy Syst.*, 2021, vol. 41, no. 1, pp. 2183–2194. doi:10.3233/JIFS-210863
10. Masum M., Shahriar H., Haddad H. M. A transfer learning with deep neural network approach for network intrusion detection. *International Journal of Intelligent Computing Research (IJICR)*, 2021, vol. 12, pp. 087–1095. doi:10.20533/ijicr.2042.4655.2021.0132
11. Noever D. A., Noever S. E. M. Image classifiers for network intrusions. *CoRR*, 2021, arXiv:2103.07765. Available at: https://arxiv.org/abs/2103.07765 (accessed 5 April 2024).
12. Ho C. M. K., Yow K.-C., Zhu Z., Aravamuthan S. Network intrusion detection via flow-to-image conversion and vision transformer classification. *IEEE Access*, 2022, vol. 10, pp. 97780–97793. doi:10.1109/ACCESS.2022.3200034
13. Dosovitskiy A., Beyer L., Kolesnikov A., Weissenborn D., Zhai X., Unterthiner T., Dehghani M., Minderer M., Heigold G., Gelly S., Uszkoreit J., Houlsby N. An image is worth 16x16 words: Transformers for image recognition at scale. 2020, arXiv:2010.11929. Available at: https:abs/2010.11929 (accessed 5 April 2024).
14. Kim T., Pak W. Deep learning-based network intrusion detection using multiple image transformers. *Appl. Sci.*, 2023, vol. 13, no. 2754. doi:10.3390/app13052754
15. Hosler R., Sundar A., Zou X., Li F., Gao T. Unsupervised deep learning for an image based network intrusion detection system. *Proc. of 2023 IEEE Global Communications Conf.*, Kuala Lumpur, Malaysia, 2023, pp. 6825–6831. doi:10.1109/GLOBECOM54140.2023.10437636
16. Golubev S., Novikova E., Fedorchenko E. Image-based approach to intrusion detection in cyber-physical objects. *Information*, 2022, vol. 13, iss. 12, pp. 553. doi:10.3390/info13120553
17. Ghadermazi J., Shah A., Bastian N. Towards real-time network intrusion detection with image-based sequential packets representation. *IEEE Transactions on Big Data*, 2024, no. 01, pp. 1–17. doi:10.1109/TBDATA.2024.3403394
18. Zhang X., Chen J., Zhou Y., Han L., Lin J. A multiple-layer representation learning model for network-based attack detection. *IEEE Access*, 2019, vol. 7, pp. 91992–92008. doi:10.1109/ACCESS.2019.2927465
19. Yu L., Dong J., Chen L., Li M., Xu B., Li Z., Qiao L., Liu L., Zhao B., Zhang B. Pbcnn: Packet bytes-based convolutional neural network for network intrusion detection. *Computer Networks*, 2021, vol. 194, no. 108117. doi:10.1016/j.comnet.2021.108117
20. Sharafaldin I., Lashkari A. H., Ghorbani A. A. Toward generating a new intrusion detection dataset and intrusion traffic characterization. *Proc. of Intern. Conf. on Information Systems Security and Privacy*, Funchal, Madeira — Portugal, January 22–24, 2018, pp. 108–116.
21. Sun Y., Esaki H., Ochiai H. Adaptive intrusion detection in the networking of large-scale LANs with segmented federated learning. *IEEE Open Journal of the Communications Society*, 2021, vol. 2, pp. 102–112. doi:10.1109/OJCOMS.2020.3044323
22. Wu P., Guo H., Buckland R. A transfer learning approach for network intrusion detection. *Proc. of 2019 IEEE 4th Intern. Conf. on Big Data Analytics (ICBDA)*, Suzhou, China, 2019, pp. 281–285. doi:10.1109/ICBDA.2019.8713213
23. Haralick R., Shanmugam K., Dinstein I. Textural features for image classification. *IEEE TSMC*, 1973, vol. 3, iss. 6, pp. 610–621.