МЕТОДИКА ЗАЩИТЫ СИСТЕМ ОБНАРУЖЕНИЯ ВТОРЖЕНИЙ ОТ СОСТЯЗАТЕЛЬНЫХ АТАК НА ОСНОВЕ ШУМОПОДАВЛЯЮЩИХ АВТОЭНКОДЕРОВ[1]

Котенко И.В.[2], Ичетовкин Е.А.[3]

***Ключевые слова**: кибербезопасность, обнаружение вторжений, глубокое обучение, атаки на компоненты машинного обучения, автоэнкодер.*

Аннотация

Цель работы: разработка методики защиты компонентов машинного обучения систем обнаружения вторжений от состязательных атак, таких как атака на основе метода быстрого знака градиента (Fast Gradient Sign Method), с использованием подсистемы защиты на основе автоэнкодеров.

Методы исследования: использование long short-term memory (LSTM) с добавлением гауссовского шума и регуляризации. Обучение модели производится на зашумленных данных, что позволяет игнорировать искажения и выделять ключевые признаки. В качестве метрик оценки для определения эффективности модели в условиях атак использовались F-мера, точность (precision) и полнота (recall). Эксперименты проводились на трех системах обнаружения вторжений (СОВ): системе обнаружения многошаговых вторжений, системе обнаружения вторжений на основе машинного обучения и системе обнаружения вторжений на основе глубокого обучения.

Результаты исследования: предложенный метод значительно повышает устойчивость СОВ к состязательным атакам. Для системы обнаружения многошаговых вторжений F-мера увеличилась с 0.67 до 0.97, для системы обнаружения вторжений на основе машинного обучения и системы обнаружения вторжений на основе глубокого обучения — с 0.57 до 0.92. Наилучшие результаты достигнуты при конфигурации количества нейронов входного слоя LSTM1=128, выходного слоя LSTM2=64 и уровне шума ε=0.2—0.3. При этом точность и полнота также демонстрируют рост, что подтверждает эффективность метода. Использование LSTM с гауссовским шумом и регуляризацией приводит к повышению надежности классификации. Разработанная подсистема защиты позволяет использовать компоненты машинного обучения в различных системах обнаружения вторжений и позволяет обеспечить устойчивость к состязательным атакам.

DOI: 10.24412/1994-1404-2025-1-110-120

Введение

Информационные инфраструктуры критических приложений являются сложными гетерогенными системами. Они достаточно разнородны и могут состоять из совокупности различных объектов. К критическим могут быть отнесены системы, при компрометации которых будет поставлена под угрозу жизнь людей, нанесен экономический ущерб, создается угроза государству. Цифровизация таких инфраструктур сделала их подверженными не только физическим воздействиям, но и киберугрозам. Кибератаки чаще всего направленны на воздействие через вычислительные сети. Злоумышленники используют различные виды атакующих воздействий. Для защиты от такого рода атак применяют системы обнаружения вторжений (СОВ). В состав современных СОВ включаются компоненты машинного обучения (МО) для выявления ранее неизвестных типов атак и аномалий поведения сети. Такой подход характеризуется приемлемыми результатами детектирования без необходимости сложной настройки системы. Однако применяемые компоненты МО сами могут становиться потенциальными целями киберпреступников из-за подверженности особым видам атак — состязательным атакам, направленным на искажение и нарушение результатов работы

[1] Работа выполнена при частичной финансовой поддержке бюджетной темы FFZF-2025-0016.

[2] ***Котенко Игорь Витальевич,*** *заслуженный деятель науки РФ, доктор технических наук, профессор, главный научный сотрудник и руководитель лаборатории проблем компьютерной безопасности Санкт-Петербургского федерального исследовательского центра Российской академии наук (СПб ФИЦ РАН), г. Санкт-Петербург, Российская Федерация. ORCID: 0000-0001-6859-7120.*

E-mail: ivkote@comsec.spb.ru

[3] ***Ичетовкин Егор Андреевич,*** *аспирант лаборатории проблем компьютерной безопасности Санкт-Петербургского Федерального исследовательского центра Российской академии наук (СПб ФИЦ РАН), г. Санкт-Петербург, Российская Федерация.*

E-mail: ichetovkin.e@iias.spb.su

компонентов МО. При атаках отравления данных злоумышленник компрометирует обучающий набор данных, а при атаках уклонения воздействует на систему МО с помощью специально сгенерированных входных данных. Такие атаки являются существенным препятствием для применения СОВ, основанных на МО [1].

Актуальность исследования заключается в нарастающем интересе в использовании компонентов МО в системах обнаружения вторжений. Так, например, в [2] предложен подход к синтезу модели МО для обнаружения сетевых атак. Такой подход позволяет эффективно выявлять аномалии в сетевом трафике. В [3] рассмотрена методика сбора обучающих данных, защищенных от атак отравлением. Этап сбора данных является ключевым этапом для повышения точности распознавания кибератак компонентами МО. В [4] проведено сравнение СОВ на основе МО с традиционными системами, основанными на сигнатурах. В [5] предложено применение глубокого обучения для анализа сетевого трафика, что позволяет повысить точность обнаружения атак, но также увеличивает сложность модели и ее уязвимость к атакам.

Полученные результаты показали, что методы МО обладают значительным преимуществом в обнаружении неизвестных ранее типов атак по сравнению с традиционными. Однако, как отмечают авторы рассмотренных работ, классификаторы, используемые в исследованиях, оказались уязвимы к состязательным атакам. Так, в [6] и [7] показано, как киберпреступники используют слабые места классификаторов на основе МО, применяя атаки на основе FGSM (Fast Gradient Sign Method, метод быстрого знака градиента). Это давно известный тип атаки, который обычно применяют для атак на классификаторы изображений. Атакующий создает некоторое количество незаметного для человека шума на входе, что приводит к неверному функционированию компонента МО. В [8] представлен анализ существующих средств защиты от состязательных атак. Выяснилось, что существующие методики защиты, например, такие, как «типичная тренировка противника» (typical enemy training), работают неэффективно. Подобные стратегии защиты часто игнорируют динамическую, многомерную природу сетевых данных [9]. Основная выявленная проблема, высвеченная в данной работе, — отсутствие работоспособной методики защиты компонентов МО систем обнаружения вторжений от состязательных атак [10].

В качестве базы для методики защиты в настоящей работе предложено использование нейронных сетей LSTM (long short-term memory, долгая краткосрочная память). Существующие исследования показывают эффективность применения данного подхода в СОВ, но только в качестве основы для обнаружения вторжений, так как LSTM хорошо справляется с последовательными шаблонами атак [11, 12]. В настоящей работе показана эффективность использования LSTM именно для защиты компонентов МО. Кроме того, отличительной чертой работы является ее практическая значимость. Упор в исследовании делается на применение метрик, которые важны для корректной оценки работоспособности классификаторов систем обнаружения вторжений. Это такие метрики, как точность (precision), полнота (recall) и F-мера [13]. Указанный подход был использован также в работах [14] и [15], в которых решалась задача оценки применимости компонентов МО систем обнаружения вторжений при воздействии состязательных атак.

В настоящей работе представлено комплексное исследование различных моделей классификаторов компонентов МО систем обнаружения вторжений, например, OC-SVM (One-Class Support Vector Machines, метод опорных векторов для одноклассовой классификации), RF (Random Forest, случайный лес) и другие. Для всех моделей было проведено моделирование атак и показана эффективность разработанной подсистемы защиты в условиях состязательных атак.

Метрики оценки воздействия и данные для моделирования

Для определения воздействия состязательных атак на компоненты МО систем обнаружения вторжений в работе используются классические метрики, применяемые для оценки эффективности классификаторов.

Точность, с которой система идентифицирует конкретную запись *l* из множества записей *L* как являющуюся частью атаки, определяется как:

$$Precision = \frac{TP}{TP+FP}. \quad (1)$$

Доля обнаруженных вредоносных событий от всех вредоносных событий в примере $l \in L$ описывается метрикой полноты системы:

$$Recall = \frac{TP}{TP+FN}. \quad (2)$$

F-мера гармонически объединяет полноту и точность:

$$F = 2\frac{precision \times recall}{precision+recall}, \quad (3)$$

здесь TP, TN — количество правильно классифицированных положительных и отрицательных примеров, а FP, FN — количество ложноположительных и ложноотрицательных результатов [16].

Набор данных для экспериментальной части — это набор данных, разработанный Канадским институтом кибербезопасности (CICIDS). CICIDS содержит актуальный набор атак, позволяющих моделировать поведение вычислительной сети в различных состояниях. В наборе данных отдельные сетевые потоки размечены временными отрезками. CICIDS содержит данные о IP-адресах, портах, протоколах и атаках.

Набор данных использует систему B-профилей (behavioral profiles, поведенческих профилей) — моделей, которые описывают поведение в сети, для моделирования работы пользователей и генерации нейтрального фонового трафика.

Взаимодействие происходит на основе протоколов HTTP (HyperText Transfer Protocol, протокол передачи гипертекста), HTTPS (HyperText Transfer Protocol Secure, защищённый протокол передачи гипертекста), FTP (File Transfer Protocol, протокол передачи файлов), SSH (Secure Shell, безопасная оболочка), SSL (Secure Sockets Layer, уровень защищенных сокетов) и протоколов электронной почты SMTP (Simple Mail Transfer Protocol, простой протокол передачи почты), POP3 (Post Office Protocol, version 3, протокол почтового отделения, версия 3), IMAP (Internet Message Access Protocol, протокол доступа к электронной почте в Интернете).

В наборе данных присутствуют различные типы атак, такие как Heartbleed (уязвимость в SSL), веб-атаки, например, SQL-инъекции, XSS (Cross-Site Scripting, межсайтовый скриптинг), CSRF (Cross-Site Request Forgery, подделка межсайтовых запросов), ботнеты (сети заражённых компьютеров), DDoS (Distributed Denial of Service, распределённая атака на отказ в обслуживании) и другие [2, 17, 18].

Исследуемые СОВ применяют компоненты МО, основанные на различных моделях классификаторов для детектирования атак. Компоненты МО были обучены на наборе CICIDS [17].

Для эксперимента были отобраны семь различных моделей МО: Наивный Байес, DBN (Deep Belief Networks, глубокие сети доверия), QDA (Quadratic Discriminant Analysis, квадратичный дискриминантный анализ), Случайный лес, ID3 (Iterative Dichotomiser 3, итеративный дихотомизатор 3), AdaBoost (Adaptive Boosting, адаптивное бустирование), MLP (Multilayer Perceptron, многослойный перцептрон) и k-ближайших (k-Nearest Neighbors, k-ближайших соседей).

Эти модели характеризуются высокой точностью обнаружения кибератак [18—20]. Сравнительная характеристика моделей классификаторов в составе систем обнаружения вторжений с компонентами машинного обучения представлена в табл. 1.

Таблица 1

Сравнительная характеристика компонентов МО СОВ

Модель	СОВ	Показатели		
		F	Precision	Recall
OC-SVM / RF	СОВ-МВ (Система обнаружения многошаговых вторжений) [18]	0.99	99.26	98.34
RF	СОВ-МО (Система обнаружения вторжений на основе МО) [19]	0.97	98.20	96.10
DBN	СОВ-ГО (Система обнаружения вторжений на основе глубокого обучения) [20]	0.94	88.70	99.70
MLP		0.87	81.70	99.50

Все модели СОВ с компонентами МО, представленные в табл. 1, реализованы на языке Python. СОВ-МВ имеет архитектуру, включающую пять уровней: предварительная фильтрация, базовая классификация, углубленный анализ, принятие решений и обратная связь с обучением. На первом уровне происходит сбор и очистка сетевого трафика, на втором — начальная классификация с использованием простых алгоритмов МО, таких как деревья решений (Decision Trees). Для обработки аномалий применен третий уровень — глубокое обучение. На четвертом уровне модель осуществляет окончательную классификацию киберугрозы. Пятый уровень обеспечивает обновление модели через механизмы непрерывного обучения.

СОВ-МВ разработана с использованием библиотек МО: Scikit-learn, TensorFlow и PyTorch. Для обработки больших объемов данных применяется Apache Spark, сбор сетевого трафика осуществляется через Wireshark и tcpdump. Данные хранятся в распределенных базах Apache Kafka и Elasticsearch. Система интегрирована с сетевыми устройствами через RESTful API и NETCONF. Разработчики СОВ-МВ увеличили устойчивость, используя регуляризацию и адаптивное обучение. Однако авторы указывают на необходимость исследований в этой области для повышения надежности [18].

В [19] описана СОВ-МО. Она построена на моделях обучения с учителем. Использованы алгоритмы: RF, SVM и градиентное усиление (Gradient Boosting). Структура включает этапы предобработки данных, выделения признаков, обучения и тестирования. На этапе предобработки данные очищаются от шума и нормализуются для улучшения качества обучения. Для выделения признаков используются PCA (Principal Component Analysis, анализ главных компонентов). Классификатор реализован с использованием библиотек Scikit-learn и XGBoost. Для ускорения применяются Pandas и NumPy. Разработчики СОВ-МО указывают на необходимость применения подсистемы защиты компонента МО [19].

СОВ-ГО использует многослойную архитектуру. Модель DBN реализована с использованием библиотек глубокого обучения TensorFlow и Keras. Система использует RBM (Restricted Boltzmann Machines, ограниченные машины Больцмана) в качестве скрытых слоев. Первый слой инициализируется входными данными, а последующие слои соединяются друг с другом. Входные данные для последующих уровней DBN также

обновляются, возвращая значение среднеквадратичной ошибки (MSE). Авторами всех вышеописанных систем обнаружения вторжений с компонентами МО высказано опасение, что они могут быть подвержены специфическим атакам на сам классификатор [18—20].

Состязательные атаки на компоненты МО

Исследование уязвимости компонентов МО проведено в рамках состязательных атак. Такие атаки представляют собой вмешательство в работу компонента МО. Злоумышленники изменяют входные данные с целью получения неверных результатов работы классификатора [21].

Атакующий, анализируя архитектуру и алгоритмы моделей, создает состязательный набор. Основная цель атакующего воздействия — добиться неверной классификации или идентификации. В качестве состязательной атаки исследователями рассмотрен подход FGSM. Метод основан на вычислении градиентов функции потерь:

$$\eta = \varepsilon sign\,(\nabla \times J(w, x, y)), \quad (4)$$

где ε — это небольшая константа, определяющая величину возмущения, $\nabla \times J(w, x, y)$ — градиент функции потерь относительно входного образа x, а y — истинная метка [22]. В данной работе FGSM служит инструментом для исследования устойчивости моделей компонентов МО СОВ.

Ранее авторами уже было проведено исследование влияния состязательных атак на компоненты МО СОВ, представленных в таблице 1 [23, 24]. Для СОВ было проведено моделирование атак FGSM. Срезы результатов моделирования представлены в табл. 2.

Таблица 2

Воздействие FGSM-атаки на показатели классификаторов СОВ

Система обнаружения вторжений	ε	Показатели		
		F	Precision (%)	Recall (%)
СОВ-МВ	0.00	0.98	99	98
	0.15	0.74	72	76
	0.30	0.57	52	64
СОВ-МО	0.00	0.97	98	96
	0.15	0.8	92	71
	0.30	0.67	84	56
СОВ-ГО	0.00	0.94	88	99
	0.15	0.79	72	88
	0.30	0.65	60	72

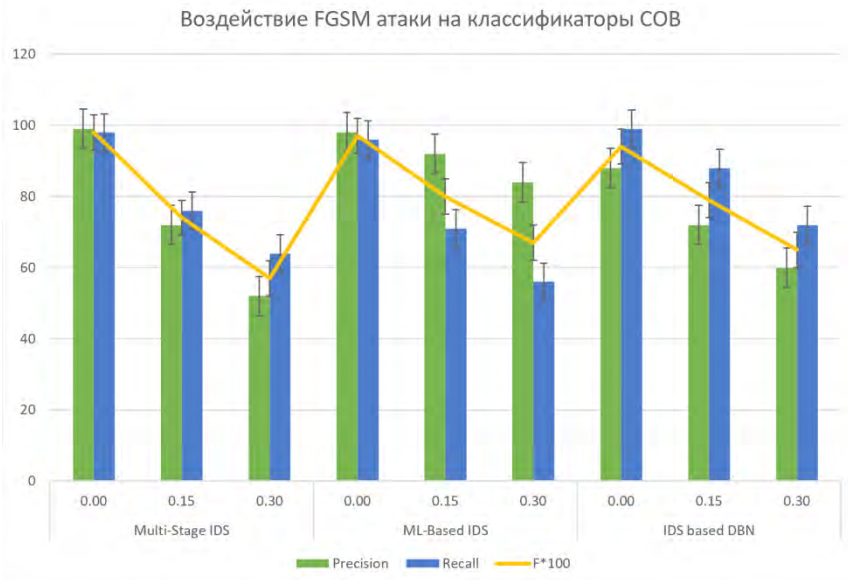

Рис. 1. Сравнительная диаграмма воздействия FGSM-атаки на показатели классификаторов СОВ

В таблице 2 сведены исследуемые показатели (1—3) ранее проведенных экспериментов по моделированию атак на компоненты МО СОВ [23]. На основе значений показателей в таблице 2 можно проанализировать тенденцию по снижению показателей классификаторов МО СОВ (рис. 1).

СОВ-МВ демонстрирует высокую точность (0.98) и полноту (0.99) при отсутствии шума (ε = 0.00), однако с увеличением ε до 0.15 и 0.30 наблюдается значительное снижение всех метрик: точность падает до 0.74 и 0.57, а полнота — до 0.76 и 0.64 соответственно. Это указывает на низкую устойчивость системы к шуму и атакам.

СОВ-МО, напротив, показывает более сбалансированные результаты. При ε = 0.00 точность (0.97) и полнота (0.96) близки к СОВ-МВ, но F-мера выше (0.97), что свидетельствует о лучшем балансе между точностью и полнотой. С увеличением ε до 0.15 точность возрастает до 0.8, а полнота снижается до 0.71, что может говорить о более адекватной классификации. При ε = 0.30 F-мера (0.67) остается выше, чем у СОВ-МВ, что подтверждает лучшую устойчивость СОВ-МО к шуму.

СОВ-ГО при ε = 0.00 демонстрирует максимальную полноту (0.99), но более низкую точность (0.94) по сравнению с другими системами. Это может указывать на склонность системы к ложным срабатываниям. С увеличением ε до 0.15 и 0.30 точность и полнота снижаются, но F-мера остается на приемлемом уровне (0.79 и 0.65 соответственно), что говорит о частичной устойчивости системы к шуму.

Вышеописанные СОВ уязвимы к шуму. Для их полноценной эксплуатации в критической инфраструктуре необходимо применение методики защиты от состязательных атак. Для этого далее будет представлена предлагаемая методика защиты компонентов МО СОВ.

Методика защиты

Существуют различные подходы к защите компонентов МО СОВ. Поскольку данные атаки так или иначе связаны с применением шума, то для защиты могут быть использованы шумоподавляющие автоэнкодеры (denoising autoencoders) [12, 25—27]. Шумоподавляющий автоэнкодер обучается на вредоносных наборах данных. В процессе обучения на вход подаются зашумленные данные, а на выходе получается значение без шума. Если x — оригинальный образ, а $\tilde{x}$ — зашумленный образ, то задача автоэнкодера заключается в минимизации разности между оригиналом и восстановленным образом:

$$L(x, \tilde{x}) = |x - \tilde{x}|^2, \quad (6)$$

где L — функция потерь (обычно используется среднеквадратичная ошибка) [25].

LSTM — это одна из архитектур шумоподавляющих автоэнкодеров. Для обучения LSTM на входные данные накладывается специально сгенерированный гауссовский шум, имитирующий атаку FGSM. Модели в таких условиях учатся отличать важные признаки исходного образа от случайного шума. После завершения обучения нейронные сети могут быть использованы для фильтрации входных состязательных образов. Такой подход позволяет повысить устойчивость компонентов МО СОВ. LSTM представляют из себя рекуррентные нейронные сети, которые способны запоминать состояния в течение длительных промежутков времени [12, 25—27].

Рассмотрим далее основные элементы LSTM: уровень (ворота) забывания (Forget Gate), вход (Input Gate), предложенное значение (Candidate Values), выход (Output Gate).

Ворота забывания определяют информацию, которую необходимо забыть:

$$f_t = \sigma(W_f \cdot [h_t - 1, x_t] + b_f), \quad (7)$$

где f_t — вектор забывания в момент t, W_f — матрица весов для забывающего слоя, b_f — вектор смещения для забывающего слоя, σ — сигмоидная функция, возвращает от 0 до 1, $[h_t - 1, x_t]$ — конкатенация предыдущего состояния $h_t - 1$ и текущего входа x_t.

Вход добавляет информацию:

$$i_t = \sigma(W_i \cdot [h_t - 1, x_t] + b_i), \quad (8)$$

где i_t — вектор входа в момент t, W_i — матрица весов для входного слоя, b_i — вектор смещения для входного слоя, выходное состояние.

Предложенное значение:

$$\mathrm{C}_t = tanh(W_\mathrm{c} \cdot [h_t - 1, x_t] + b_\mathrm{c}), \quad (9)$$

где C_t — кандидатные значения в момент t, W_c — матрица весов для кандидатных значений, b_c — вектор смещения для кандидатных значений, $tanh$ — гиперболический тангенс, значения от -1 до 1.

Обновление состояния:

$$\mathrm{C}_t = f_t\, \mathrm{C}_{t-1} + i_t\, \widetilde{\mathrm{C}_t}, \quad (10)$$

где C_t — новое состояние в момент t, C_{t-1} — предыдущее состояние на шаге t – 1, $\widetilde{\mathrm{C}_t}$ — вектор новых значений.

Выход:

$$o_t = \sigma(W_o \cdot [h_t - 1, x_t] + b_o), \quad (11)$$

где o_t — вектор выхода в момент t, W_o — матрица весов для выходного слоя, b_o — вектор смещения для выходного слоя.

Выходное состояние LSTM:

$$h_t = o_t\, tanh(\mathrm{C}_t), \quad (12)$$

где h_t — скрытое состояние (выход) в момент t.

Схема, описывающая реализацию подсистемы защиты на основе автоэнкодера с двумя LSTM слоями (входящий и выходящий), представлена на рис. 2.

Методика защиты компонентов МО СОВ с использованием подсистемы защиты включает в себя следующие блоки: загрузка и предобработка данных; создание модели LSTM; реализация атаки FGSM; обучение модели с защитой от атак; тестирование модели.

Обучение модели проводится с использованием средства оптимизации Adam. Для предлагаемой подсистемы защиты применяются следующие параметры: сгенерированный гауссовский уровень шума (Noise), количество нейронов входного (LSTM1) и выходного слоев (LSTM2). Такая настройка позволит оценить эффективность методики, реализованной в подсистеме защиты от атак уклонением на компоненты МО СОВ. Это нивелирует последствия FGSM-атаки.

Рис. 2. Подсистема защиты компонентов МО СОВ

Анализ эффективности методики

Для оценки эффективности подсистемы защиты используем метрики (1—3). Модели классификаторов, которые исследованы, представлены в таблице 1. Сетевые атаки смоделированы проверочной выборкой из CICIDS. В качестве атак на компоненты МО СОВ применяется FGSM с параметром $\varepsilon = 0.3$ (4). Подсистема защиты реализована на базе LSTM.

В эксперименте использования подсистемы защиты для СОВ-МВ наблюдаются признаки линейной зависимости показателей (1—3) на тестируемом участке. Можно предположить, что локальное насыщение наступает в конце при количестве нейронов входного слоя *LSTM1* = 128, выходного *LSTM2* = 64, уровня гауссовского шума *Noise* = 0,3. Результаты восстановления параметров детектирования: *F* = 0.92, *Precision* = 88, *Recall* = 96.

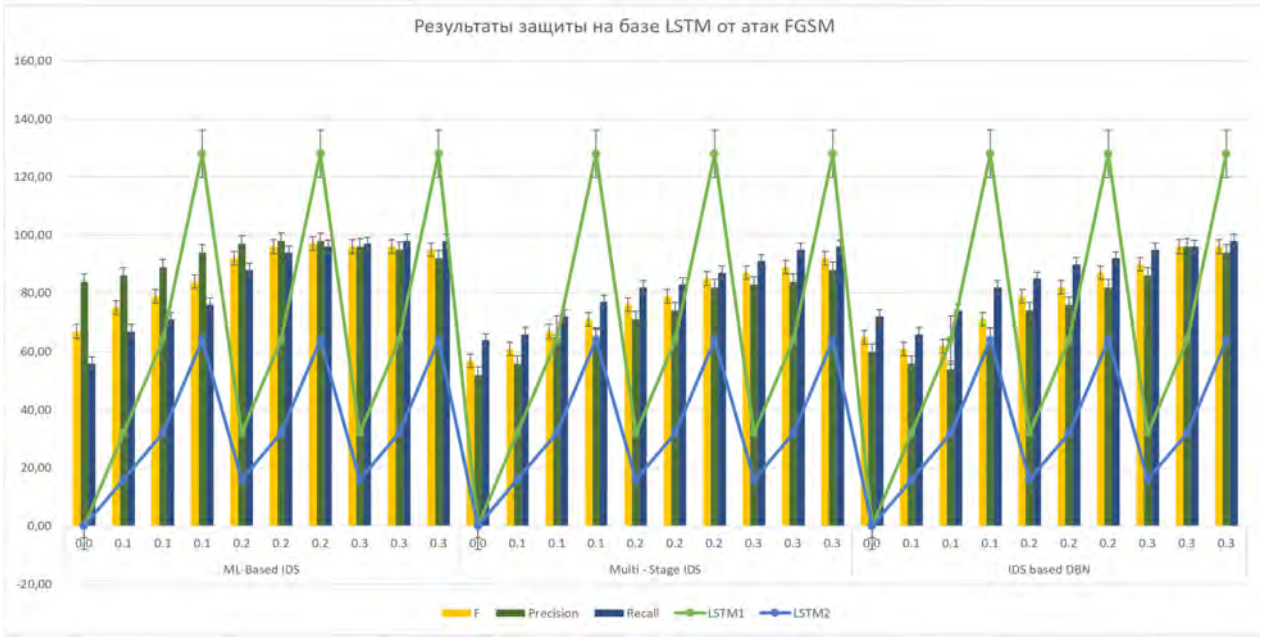

Рис. 3. Результаты работы подсистемы защиты

Таблица 3

Воздействие FGSM-атаки на показатели классификаторов СОВ

СОВ	FGSM	Подсистема защиты			Показатели		
					F*100	Precision (%)	Recall (%)
	ε	Noise	LSTM1	LSTM2			
СОВ-МВ	0.30	0.00	0	0	67,00	84,00	56,00
		0.10	32	16	75,00	86,00	67,00
		0.10	64	32	79,00	89,00	71,00
		0.10	128	64	84,00	94,00	76,00
		0.20	32	16	92,00	97,00	88,00
		0.20	64	32	96,00	98,00	94,00
		0.20	128	64	97,00	98,00	96,00
		0.30	32	16	96,00	96,00	97,00
		0.30	64	32	96,00	95,00	98,00
		0.30	128	64	95,00	92,00	98,00
СОВ-МО		0.00	0	0	57,00	52,00	64,00
		0.10	32	16	61,00	56,00	66,00
		0.10	64	32	67,00	63,00	72,00
		0.10	128	64	71,00	65,00	77,00
		0.20	32	16	76,00	71,00	82,00
		0.20	64	32	79,00	74,00	83,00
		0.20	128	64	85,00	82,00	87,00
		0.30	32	16	87,00	83,00	91,00
		0.30	64	32	89,00	84,00	95,00
		0.30	128	64	92,00	88,00	96,00
СОВ-ГО		0.00	0	0	57,00	52,00	64,00
		0.10	32	16	61,00	56,00	66,00
		0.10	64	32	67,00	63,00	72,00
		0.10	128	64	71,00	65,00	77,00
		0.20	32	16	76,00	71,00	82,00
		0.20	64	32	79,00	74,00	83,00
		0.20	128	64	85,00	82,00	87,00
		0.30	32	16	87,00	83,00	91,00
		0.30	64	32	89,00	84,00	95,00
		0.30	128	64	92,00	88,00	96,00

Для СОВ-МО удалось практически полностью восстановить работоспособность системы. Пределом восстановления стали параметры *LSTM1* = 32, *LSTM2* = 16, *Noise* = 0,3. При таких параметрах удалось добиться следующих результатов работы системы: *F* = 0.96, *Precision* = 96, *Recall* = 97. Разница между исходными параметрами незначительная.

В эксперименте с СОВ-ГО наблюдаются потенциально интересные явления для гармонической метрики *F*, которая в конце сближается с *Recall*. Насыщение проис-

ходит при параметрах, близких к *LSTM1* = 64, *LSTM2* = 32, *Noise* = 0.3. Получены результаты *F* = 0.96, *Precision* = 94, *Recall* = 98. Параметры восстановлены до приемлемого уровня. Полученные параметры работы подсистемы защиты представлены в табл. 3 выше.

Таким образом, в экспериментах были исследованы системы обнаружения вторжений с компонентами МО, описанными в [18—20]. Результаты экспериментов показали, что применение предложенной методики защиты позволяет защитить компоненты МО СОВ от атак уклонением и восстановить метрики *F, Precision, Recall* фактически до исходных параметров, что изображено на рис. 3 выше.

Предложенный подход демонстрирует достаточно высокую эффективность защиты компонентов машинного обучения систем обнаружения вторжений от состязательных атак, таких как FGSM. В отличие от традиционных методов, которые фокусируются на повышении точности классификации, данный подход решает проблему устойчивости моделей к атакам. Использование LSTM с добавлением гауссовского шума ($\varepsilon = 0.1$—0.3) и регуляризации (Dropout, L2) позволяет модели эффективно классифицировать атаки и противостоять искажениям входных данных. Это подтверждается ростом *F*-меры с 0.67 до 0.97 для СОВ-МВ, с 0.57 до 0.92 для СОВ-МО и СОВ-ГО.

Заключение

В статье представлена методика защиты компонентов МО СОВ критических инфраструктур. В качестве основного инструмента защиты предлагается применение подсистемы защиты на базе Long Short-Term Memory. Проведен обзор релевантных работ и описан математический аппарат защиты. Проведено моделирование атак уклонением Fast Gradient Sign Method с последующей защитой. Предложенная методика защиты реализована на языке Python. Методика показала возможность восстановления метрик детектирования *F, precision, recall* к значениям, близким к исходным. Полученные показатели эффективности превосходят известные методы, такие как TIDCS (A Dynamic Intrusion Detection and Classification System Based on Feature Selection, динамическая система обнаружения и классификации вторжений на основе выбора признаков) [2], показатели которого для *F*-меры 0.85. В предложенной реализации удалось достичь значений 0.97 для СОВ-МВ и 0.92 для СОВ-МО и СОВ-ГО.

Следует отметить, что исследование применимости LSTM для сетевых данных еще не завершено, и полученные результаты являются базисом для последующих исследований. Существующие исследования, например, [13] и [21], рассматривают LSTM только в приложениях компьютерного зрения и промышленных системах. Обзорные работы, например, [28], дают только общее представление о методах защиты, без раскрытия конкретных решений. Новизна предложенной методики заключается в комбинировании LSTM, состязательного обучения и регуляризации. Полученные результаты направлены на повышение устойчивости СОВ к состязательным атакам.

Направление дальнейшего исследования видится в разработке методики защиты компонентов МО СОВ более широкого спектра классификаторов.

Рецензент: ***Лаута Олег Сергеевич,*** *доктор технических наук, профессор кафедры комплексного обеспечения информационной безопасности Государственного университета морского и речного флота имени адмирала С.О. Макарова, г. Санкт-Петербург, Россия.*
E-mail: laos-82@yandex.ru

Литература

1. Ковцур М.М., Кириллов Д.И., Михайлова А.В., Потемкин П.А. Разработка методики внедрения машинного обучения для повышения информационной безопасности web-приложения // Техника средств связи. 2020. № 4 (152). С. 74—86.
2. Chkirbene Z., Erbad A., Hamila R., Mohamed A., Guizani M., Hamdi M. TIDCS: A Dynamic Intrusion Detection and Classification System Based Feature Selection. IEEE Access. 2020. Vol. 8. P. 95864—95877. DOI: 10.1109/ACCESS.2020.2994959 .
3. Гетьман А.И., Горюнов М.Н., Мацкевич А.Г., Рыболовлев Д.А. Методика сбора обучающего набора данных для модели обнаружения компьютерных атак // Труды ИСП РАН. 2021. Т. 33. Вып. 5. С. 83—104. DOI: 10.15514/ISPRAS-2021-33(5)-5 .
4. Гетьман А.И., Горюнов М.Н., Мацкевич А.Г., Рыболовлев Д.А. Сравнение системы обнаружения вторжений на основе машинного обучения с сигнатурными средствами защиты информации // Труды ИСП РАН. 2022. Т. 34. Вып. 5. С. 111—126. DOI: 10.15514/ISPRAS-2022-34(5)-7 .
5. Гетьман А.И., Горюнов М.Н., Мацкевич А.Г., Никольская А.Г., Рыболовлев Д.А. Состязательные атаки против системы обнаружения вторжений, основанной на применении методов машинного обучения // Проблемы информационной безопасности. Компьютерные системы. 2023. № 4. С. 156—190. DOI: 10.48612/jisp/eatr-5pxb-akt8 .

6. Alhajjar E., Maxwell P., Bastian N. Adversarial Machine Learning in Network Intrusion Detection Systems. Expert Systems with Applications. 2021. Vol. 186. P. 115782. DOI: 10.1016/j.eswa.2021.115782 .
7. Alotaibi A., Rassam M.A. Adversarial Machine Learning Attacks Against Intrusion Detection Systems: A Survey on Strategies and Defense. Future Internet. 2023. Vol. 15. No. 2. P. 62. DOI: 10.3390/fi15020062 .
8. Apruzzese G., Andreolini M., Ferretti L., Marchetti M., Colajanni M. Modeling Realistic Adversarial Attacks Against Network Intrusion Detection Systems. Digital Threats: Research and Practice. 2022. Vol. 3. No. 3. P. 1–19. DOI: 10.1145/3530870 .
9. Гетьман А.И., Горюнов М.Н., Мацкевич А.Г., Рыболовлев Д.А., Никольская А.Г. Применение глубокого обучения для обнаружения компьютерных атак в сетевом трафике // Труды ИСП РАН. 2023. Т. 35. Вып. 4. С. 65—92. DOI: 10.15514/ISPRAS-2023-35(4)-3 .
10. Котенко И.В., Саенко И.Б., Лаута О.С., Васильев Н.А., Садовников В.Е. Атаки и методы защиты в системах машинного обучения: анализ современных исследований // Вопросы кибербезопасности. 2024. № 1 (59). С. 24—37. DOI: 10.21681/2311-2024-1-24-37 .
11. Ravi V., Chaganti R., Alazab M. Recurrent Deep Learning-Based Feature Fusion Ensemble Meta-Classifier Approach for Intelligent Network Intrusion Detection System. Computers and Electrical Engineering. 2022. Vol. 102. P. 108156. DOI: 10.1016/j.compeleceng.2022.108156 .
12. Nazir A. et al. A Deep Learning-Based Novel Hybrid CNN-LSTM Architecture for Efficient Detection of Threats in the IoT Ecosystem // Ain Shams Engineering Journal. 2024. P. 102777. DOI: 10.1016/j.asej.2024.102777 .
13. Akhtar N., Mian A., Kardan N., Shah M. Advances in Adversarial Attacks and Defenses in Computer Vision: A Survey. IEEE Access. 2021. Vol. 9. P. 155161–155196. DOI: 10.1109/ACCESS.2021.3127960 .
14. Alahmed S., Alasad Q., Hammood M.M., Yuan J.-S., Alawad M. Mitigation of Black-Box Attacks on Intrusion Detection Systems-Based ML. Computers. 2022. Vol. 11. No. 7. P. 115. DOI: 10.3390/computers11070115 .
15. Mohammadian H., Ghorbani A.A., Lashkari A.H. A Gradient-Based Approach for Adversarial Attack on Deep Learning-Based Network Intrusion Detection Systems. Applied Soft Computing. 2023. Vol. 137. P. 110173. DOI: 10.1016/j.asoc.2023.110173 .
16. Ahmad Z., Khan A.S., Shiang C.W., Abdullah J., Ahmad F. Network Intrusion Detection System: A Systematic Study of Machine Learning and Deep Learning Approaches. Transactions on Emerging Telecommunications Technologies. 2021. Vol. 32. No. 1. P. e4150. DOI: 10.1002/ett.4150 .
17. Kurniabudi, Stiawan D., Darmawijoyo, Idris M.Y. Bin, Bamhdi A.M., Budiarto R. CICIDS-2017 Dataset Feature Analysis With Information Gain for Anomaly Detection. IEEE Access. 2020. Vol. 8. P. 132911–132921. DOI: 10.1109/ACCESS.2020.3009843 .
18. Verkerken M. et al. A Novel Multi-Stage Approach for Hierarchical Intrusion Detection. IEEE Transactions on Network and Service Management. 2023. Vol. 20. No. 3. P. 3915–3929. DOI: 10.1109/TNSM.2023.3259474 .
19. Горюнов М.Н., Мацкевич А.Г., Рыболовлев Д.А. Синтез модели машинного обучения для обнаружения компьютерных атак на основе набора данных CICIDS2017 // Труды ИСП РАН. 2020. Т. 32. Вып. 5. С. 81—94. DOI: 10.15514/ISPRAS-2020-32(5)-6 .
20. Belarbi O., Khan A., Carnelli P., Spyridopoulos T. An Intrusion Detection System Based on Deep Belief Networks // The 4th International Conference on Science of Cyber Security (SciSec 2022). Springer International Publishing, Cham, 2022. P. 377–392. DOI: 10.48550/arXiv.2207.02117 .
21. Anthi E., Williams L., Rhode M., Burnap P., Wedgbury A. Adversarial Attacks on Machine Learning Cybersecurity Defences in Industrial Control Systems. Journal of Information Security and Applications. 2021. Vol. 58. P. 102717. DOI: 10.1016/j.jisa.2021.102717 .
22. Aldweesh A., Derhab A., Emam A.Z. Deep Learning Approaches for Anomaly-Based Intrusion Detection Systems: A Survey, Taxonomy, and Open Issues. Knowledge-Based Systems. 2020. Vol. 189. P. 105124. DOI: 10.1016/j.knosys.2019.105124 .
23. Ichetovkin E., Kotenko I. Modeling Poisoning Attacks Against Machine Learning Components of Intrusion Detection Systems. 2024 IEEE 25th International Conference of Young Professionals in Electron Devices and Materials (EDM). Altai, Russian Federation, 2024. P. 1850–1855. DOI: 10.1109/EDM61683.2024.10615198 .
24. Ichetovkin E., Kotenko I. Modeling Attacks on Machine Learning Components of Intrusion Detection Systems. 2024 International Russian Smart Industry Conference (SmartIndustryCon). Sochi, Russian Federation, 2024. P. 261–266. DOI: 10.1109/SmartIndustryCon61328.2024.10515506 .
25. Laghrissi F.E., Douzi S., Douzi K., Hssina B. Intrusion Detection Systems Using Long Short-Term Memory (LSTM) // Journal of Big Data. 2021. Vol. 8. No. 1. P. 65. DOI: 10.1186/s40537-021-00453-7 .
26. Ayub M.A., Johnson W.A., Talbert D.A., Siraj A. Model Evasion Attack on Intrusion Detection Systems Using Adversarial Machine Learning. 2020 54th Annual Conference on Information Sciences and Systems (CISS). 2020. P. 1–6. DOI: 10.1109/CISS48834.2020.1570617295 .
27. Chou D., Jiang M. A Survey on Data-Driven Network Intrusion Detection. ACM Computing Surveys (CSUR). 2021. Vol. 54. No. 9. P. 1–36. DOI: 10.1145/3477132 .
28. Jmila H., Khedher M.I. Adversarial Machine Learning for Network Intrusion Detection: A Comparative Study. Computer Networks. 2022. Vol. 214. P. 109073. DOI: 10.1016/j.comnet.2022.109073 .

SECTION:
INFORMATION AND AUTOMATED SYSTEMS AND NETWORKS

A TECHNIQUE FOR PROTECTING INTRUSION DETECTION SYSTEMS AGAINST ADVERSARIAL ATTACKS BASED ON DENOISING AUTOENCODERS

Igor' Kotenko, *Honoured Figure of Science of the Russian Federation, Dr.Sc. (Technology), Professor, Principal Researcher and Head of the Laboratory for Computer Security Problems of the Saint Petersburg Federal Research Centre of the Russian Academy of Sciences, Saint Petersburg, Russian Federation. ORCID: 0000-0001-6859-7120.*
E-mail: ivkote@comsec.spb.ru

Egor Ichetovkin, *Ph.D. student at the Laboratory for Computer Security Problems of the Saint Petersburg Federal Research Centre of the Russian Academy of Sciences, Saint Petersburg, Russian Federation. E-mail: ichetovkin.e@iias.spb.su*

Keywords: *cyber security, intrusion detection, deep learning, attacks against machine learning components, autoencoder.*

Abstract

Purpose of the work: development of a technique to protect machine learning components of intrusion detection systems from adversarial attacks such as Fast Gradient Sign Method using an autoencoder based defense subsystem.

Methods used in the study: application of long short-term memory (LSTM) with the addition of Gaussian noise and regularisation. The model is trained on noisy data, which allows it to ignore distortions and highlight key features. F-measure, precision and recall were used as evaluation metrics to assess the performance of the model under attack. Experiments were conducted on three intrusion detection systems (IDS): Multi-Stage IDS, Machine Learning-Based IDS and Deep Learning-Based IDS.

Study findings: the technique significantly improves the robustness of IDSs to adversarial attacks. For Multi-Stage IDS, the F-measure increased from 0.67 to 0.97, for Machine Learning-Based IDS and Deep Learning-Based IDS from 0.57 to 0.92. The best results are achieved when the configuration of the number of neurons of the input layer LSTM1=128, the output layer LSTM2=64 and the noise level ε=0.2-0.3. Precision and recall also show an increase, which confirms the effectiveness of the method. The use of LSNM with Gaussian noise and regularisation improves the reliability of classification. The technique provides robustness against adversarial attacks. The developed defense subsystem allows the use of machine learning components in different intrusion detection systems.

References

1. Kovtsur M.M., Kirillov D.I., Mikhailova A.V., Potemkin P.A. Razrabotka metodiki vnedreniia mashinnogo obucheniia dlia povysheniia informatsionnoi bezopasnosti web-prilozheniia. Tekhnika sredstv sviazi. 2020. No. 4 (152). Pp. 74–86.
2. Chkirbene Z., Erbad A., Hamila R., Mohamed A., Guizani M., Hamdi M. TIDCS: A Dynamic Intrusion Detection and Classification System Based Feature Selection. IEEE Access. 2020. Vol. 8. P. 95864–95877. DOI: 10.1109/ACCESS.2020.2994959 .
3. Get'man A.I., Goriunov M.N., Matskevich A.G., Rybolovlev D.A. Metodika sbora obuchaiushchego nabora dannykh dlia modeli obnaruzheniia komp'iuternykh atak. Trudy ISP RAN. 2021. T. 33. Vyp. 5. Pp. 83–104. DOI: 10.15514/ISPRAS-2021-33(5)-5 .
4. Get'man A.I., Goriunov M.N., Matskevich A.G., Rybolovlev D.A. Sravnenie sistemy obnaruzheniia vtorzhenii na osnove mashinnogo obucheniia s signaturnymi sredstvami zashchity informatsii. Trudy ISP RAN. 2022. T. 34. Vyp. 5. Pp. 111–126. DOI: 10.15514/ISPRAS-2022-34(5)-7 .
5. Get'man A.I., Goriunov M.N., Matskevich A.G., Nikol'skaia A.G., Rybolovlev D.A. Sostiazatel'nye ataki protiv sistemy obnaruzheniia vtorzhenii, osnovannoi na primenenii metodov mashinnogo obucheniia. Problemy informatsionnoi bezopasnosti. Komp'iuternye sistemy. 2023. No. 4. Pp. 156–190. DOI: 10.48612/jisp/eatr-5pxb-akt8 .
6. Alhajjar E., Maxwell P., Bastian N. Adversarial Machine Learning in Network Intrusion Detection Systems. Expert Systems with Applications. 2021. Vol. 186. P. 115782. DOI: 10.1016/j.eswa.2021.115782 .
7. Alotaibi A., Rassam M.A. Adversarial Machine Learning Attacks Against Intrusion Detection Systems: A Survey on Strategies and Defense. Future Internet. 2023. Vol. 15. No. 2. P. 62. DOI: 10.3390/fi15020062 .

8. Apruzzese G., Andreolini M., Ferretti L., Marchetti M., Colajanni M. Modeling Realistic Adversarial Attacks Against Network Intrusion Detection Systems. Digital Threats: Research and Practice. 2022. Vol. 3. No. 3. P. 1–19. DOI: 10.1145/3530870 .
9. Get'man A.I., Goriunov M.N., Matskevich A.G., Rybolovlev D.A., Nikol'skaia A.G. Primenenie glubokogo obucheniia dlia obnaruzheniia komp'iuternykh atak v setevom trafike. Trudy ISP RAN. 2023. T. 35. Vyp. 4. Pp. 65–92. DOI: 10.15514/ISPRAS-2023-35(4)-3 .
10. Kotenko I.V., Saenko I.B., Lauta O.S., Vasil'ev N.A., Sadovnikov V.E. Ataki i metody zashchity v sistemakh mashinnogo obucheniia: analiz sovremennykh issledovanii. Voprosy kiberbezopasnosti. 2024. No. 1 (59). Pp. 24–37. DOI: 10.21681/2311-2024-1-24-37 .
11. Ravi V., Chaganti R., Alazab M. Recurrent Deep Learning-Based Feature Fusion Ensemble Meta-Classifier Approach for Intelligent Network Intrusion Detection System. Computers and Electrical Engineering. 2022. Vol. 102. P. 108156. DOI: 10.1016/j.compeleceng.2022.108156 .
12. Nazir A. et al. A Deep Learning-Based Novel Hybrid CNN-LSTM Architecture for Efficient Detection of Threats in the IoT Ecosystem. Ain Shams Engineering Journal. 2024. P. 102777. DOI: 10.1016/j.asej.2024.102777 .
13. Akhtar N., Mian A., Kardan N., Shah M. Advances in Adversarial Attacks and Defenses in Computer Vision: A Survey. IEEE Access. 2021. Vol. 9. P. 155161–155196. DOI: 10.1109/ACCESS.2021.3127960 .
14. Alahmed S., Alasad Q., Hammood M.M., Yuan J.-S., Alawad M. Mitigation of Black-Box Attacks on Intrusion Detection Systems-Based ML. Computers. 2022. Vol. 11. No. 7. P. 115. DOI: 10.3390/computers11070115 .
15. Mohammadian H., Ghorbani A.A., Lashkari A.H. A Gradient-Based Approach for Adversarial Attack on Deep Learning-Based Network Intrusion Detection Systems. Applied Soft Computing. 2023. Vol. 137. P. 110173. DOI: 10.1016/j.asoc.2023.110173 .
16. Ahmad Z., Khan A.S., Shiang C.W., Abdullah J., Ahmad F. Network Intrusion Detection System: A Systematic Study of Machine Learning and Deep Learning Approaches. Transactions on Emerging Telecommunications Technologies. 2021. Vol. 32. No. 1. P. e4150. DOI: 10.1002/ett.4150 .
17. Kurniabudi, Stiawan D., Darmawijoyo, Idris M.Y. Bin, Bamhdi A.M., Budiarto R. CICIDS-2017 Dataset Feature Analysis With Information Gain for Anomaly Detection. IEEE Access. 2020. Vol. 8. P. 132911–132921. DOI: 10.1109/ACCESS.2020.3009843 .
18. Verkerken M. et al. A Novel Multi-Stage Approach for Hierarchical Intrusion Detection. IEEE Transactions on Network and Service Management. 2023. Vol. 20. No. 3. P. 3915–3929. DOI: 10.1109/TNSM.2023.3259474 .
19. Goriunov M.N., Matskevich A.G., Rybolovlev D.A. Sintez modeli mashinnogo obucheniia dlia obnaruzheniia komp'iuternykh atak na osnove nabora dannykh CICIDS2017. Trudy ISP RAN. 2020. T. 32. Vyp. 5. Pp. 81–94. DOI: 10.15514/ISPRAS-2020-32(5)-6 .
20. Belarbi O., Khan A., Carnelli P., Spyridopoulos T. An Intrusion Detection System Based on Deep Belief Networks // The 4th International Conference on Science of Cyber Security (SciSec 2022). Springer International Publishing, Cham, 2022. P. 377–392. DOI: 10.48550/arXiv.2207.02117 .
21. Anthi E., Williams L., Rhode M., Burnap P., Wedgbury A. Adversarial Attacks on Machine Learning Cybersecurity Defences in Industrial Control Systems. Journal of Information Security and Applications. 2021. Vol. 58. P. 102717. DOI: 10.1016/j.jisa.2021.102717 .
22. Aldweesh A., Derhab A., Emam A.Z. Deep Learning Approaches for Anomaly-Based Intrusion Detection Systems: A Survey, Taxonomy, and Open Issues. Knowledge-Based Systems. 2020. Vol. 189. P. 105124. DOI: 10.1016/j.knosys.2019.105124 .
23. Ichetovkin E., Kotenko I. Modeling Poisoning Attacks Against Machine Learning Components of Intrusion Detection Systems. 2024 IEEE 25th International Conference of Young Professionals in Electron Devices and Materials (EDM). Altai, Russian Federation, 2024. P. 1850–1855. DOI: 10.1109/EDM61683.2024.10615198 .
24. Ichetovkin E., Kotenko I. Modeling Attacks on Machine Learning Components of Intrusion Detection Systems. 2024 International Russian Smart Industry Conference (SmartIndustryCon). Sochi, Russian Federation, 2024. P. 261–266. DOI: 10.1109/SmartIndustryCon61328.2024.10515506 .
25. Laghrissi F.E., Douzi S., Douzi K., Hssina B. Intrusion Detection Systems Using Long Short-Term Memory (LSTM) // Journal of Big Data. 2021. Vol. 8. No. 1. P. 65. DOI: 10.1186/s40537-021-00453-7 .
26. Ayub M.A., Johnson W.A., Talbert D.A., Siraj A. Model Evasion Attack on Intrusion Detection Systems Using Adversarial Machine Learning. 2020 54th Annual Conference on Information Sciences and Systems (CISS). 2020. P. 1–6. DOI: 10.1109/CISS48834.2020.1570617295 .
27. Chou D., Jiang M. A Survey on Data-Driven Network Intrusion Detection. ACM Computing Surveys (CSUR). 2021. Vol. 54. No. 9. P. 1–36. DOI: 10.1145/3477132 .
28. Jmila H., Khedher M.I. Adversarial Machine Learning for Network Intrusion Detection: A Comparative Study. Computer Networks. 2022. Vol. 214. P. 109073. DOI: 10.1016/j.comnet.2022.109073 .