Научная статья
УДК 004
https://doi.org/10.24143/2072-9502-2025-2-76-87
EDN MZOXTJ

Исследование устойчивости систем обнаружения вторжений с компонентами машинного обучения к состязательным атакам

Егор Андреевич Ичетовкин

Санкт-Петербургский Федеральный исследовательский центр Российской академии наук,
Санкт-Петербург, Россия, ichetovkin.e@iias.spb.su

Аннотация. В условиях стремительного развития киберугроз современные системы обнаружения вторжений становятся ключевым элементом защиты информационной инфраструктуры. Их задача – не только выявлять известные атаки, но и обнаруживать новые, ранее не встречавшиеся угрозы, включая сложные целевые воздействия. Однако сами алгоритмы машинного обучения (МО) могут становиться объектами атак, направленных на их обход или манипуляцию результатами детектирования. Проводится детальное исследование уязвимости моделей МО к целенаправленным вредоносным воздействиям, включая атаки уклонения, когда злоумышленник намеренно модифицирует входные данные, чтобы обойти защитные механизмы. Методология исследования включает анализ существующих защитных стратегий, а также моделирование различных сценариев атак для оценки устойчивости алгоритмов. В качестве критериев эффективности применяются классические метрики: точность, полнота и *F-мера*. Показатели позволяют оценить как качество детектирования, так и степень деградации модели под воздействием атак. Практическая ценность исследования заключается в проведении комплексного сравнительного анализа устойчивости различных моделей МО, включая популярные решения, используемые в промышленных системах безопасности. Впервые тестируются несколько типов классификаторов (например, одноклассовые МО векторов, случайный лес и глубокие нейронные сети) в условиях целенаправленных атак, имитирующих действия продвинутого злоумышленника, атакующего компоненты МО систем обнаружения вторжений сложной инфраструктуры. Результаты экспериментальной оценки оказались тревожными – ни одна из рассмотренных моделей не продемонстрировала достаточной устойчивости к исследуемым атакам. Это указывает на системную уязвимость современных методов МО, применяемых в кибербезопасности, и подчеркивает необходимость разработки новых защитных механизмов, устойчивых к целенаправленному противодействию. Полученные данные могут быть использованы для совершенствования алгоритмов обнаружения вторжений и создания более надежных систем защиты.

Ключевые слова: кибербезопасность, системы обнаружения вторжений, компоненты машинного обучения, атаки уклонением, устойчивость защиты

Для цитирования: *Ичетовкин Е. А.* Исследование устойчивости систем обнаружения вторжений с компонентами машинного обучения к состязательным атакам // Вестник Астраханского государственного технического университета. Серия: Управление, вычислительная техника и информатика. 2025. № 2. С. 76–87. https://doi.org/10.24143/2072-9502-2025-2-76-87. EDN MZOXTJ.

Original article

Investigating the resistance of intrusion detection systems with machine learning components to adversarial attacks

Egor A. Ichetovkin

St. Petersburg Federal Research Center of the Russian Academy of Sciences,
Saint Petersburg, Russia, ichetovkin.e@iias.spb.su

Abstract. With the rapid development of cyber threats, modern intrusion detection systems are becoming a key element of information infrastructure protection. Their task is not only to identify known attacks, but also to detect new, previously unknown threats, including complex targeted attacks. However, machine learning (ML) algorithms themselves can become targets of attacks aimed at bypassing them and manipulating detection results. A detailed study is

being conducted on the vulnerability of ML models to targeted malicious influences, including evasion attacks, when an attacker intentionally modifies input data in order to circumvent security mechanisms. The research methodology includes an analysis of existing defensive strategies, as well as modeling various attack scenarios to assess the resilience of algorithms. Classical metrics are used as performance criteria: accuracy, completeness, and F-measure. The indicators allow us to assess both the quality of detection and the degree of degradation of the model under the influence of attacks. The practical value of the research lies in conducting a comprehensive comparative analysis of the stability of various ML models, including popular solutions used in industrial security systems. For the first time, several types of classifiers are being tested (for example, single-class vector ML, random forests, and deep neural networks) under targeted attacks that simulate the actions of an advanced attacker attacking ML components of intrusion detection systems of complex infrastructure. The results of the experimental evaluation turned out to be alarming – none of the considered models demonstrated sufficient resistance to the attacks under study. This indicates the systemic vulnerability of modern defense ML methods used in cybersecurity and underlines the need to develop new defense mechanisms that are resistant to targeted counteraction. The data obtained can be used to improve intrusion detection algorithms and create more reliable protection systems.

Keywords: cybersecurity, intrusion detection systems, machine learning components, evasion attacks, defense resilience

For citation: Ichetovkin E. A. Investigating the resistance of intrusion detection systems with machine learning components to adversarial attacks. *Vestnik of Astrakhan State Technical University. Series: Management, computer science and informatics*. 2025;2:76-87. (In Russ.). https://doi.org/10.24143/2072-9502-2025-2-76-87. EDN MZOXTJ.

Введение

В последние годы сложные ИТ-инфраструктуры сталкиваются с беспрецедентным давлением со стороны злоумышленников, чему способствует стремительная цифровая трансформация процессов [1]. Как показывают исследования, экспоненциальное увеличение количества интеллектуальных устройств в рамках концепции Интернета вещей не только создает новые векторы атак, но и многократно усиливает потенциальный ущерб от успешных кибернетических инцидентов [2].

Парадоксально, но современные инструменты безопасности становятся заложниками собственной архитектуры – их ресурсы часто оказываются недостаточными для обработки растущего потока угроз, что приводит к запаздыванию реакции на инциденты [3]. Сложившаяся ситуация напоминает асимметричную войну, где злоумышленники традиционно обладают инициативой, вынуждая специалистов по защите информации постоянно совершенствовать оборонительные механизмы.

Особую роль в этой борьбе играют системы обнаружения вторжений (СОВ, Intrusion Detection Systems – IDS), демонстрирующие эффективность против широкого спектра угроз – от шаблонных атак до сложных целевых компрометаций (рис. 1).

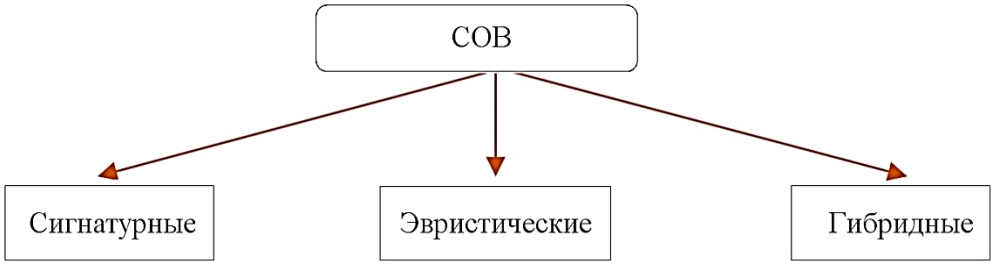

Рис. 1. Разновидности систем обнаружения вторжений

Fig. 1. Types of intrusion detection systems

Современные СОВ применяют разнообразные методики выявления киберугроз. Наиболее традиционный подход – сигнатурный анализ, при котором система сверяет сетевую активность с базой известных шаблонов атак (так называемых «цифровых отпечатков»). Хотя такой метод демонстрирует высокую точность при идентификации уже изученных угроз и минимальный уровень ложных тревог, его принципиальным ограничением остается неспособность распознавать ранее неизвестные атаки, включая угрозы нулевого дня [4].

Альтернативой выступает технология обнаружения аномалий, построенная на принципах машинного обучения (МО). В этом случае система сначала обучается на «нормальном» поведении системы, а затем флагирует любые значительные отклонения от этого эталона. Такой эвристический подход, по данным исследований [5], действительно позволяет выявлять новые виды угроз, од-

нако требует тщательной настройки для минимизации ложноположительных срабатываний.

Практический опыт показывает, что оптимальным решением часто становится гибридная архитектура, сочетающая преимущества обоих методов. В таких системах сигнатурный анализ дополняется возможностями МО, при этом выявленные аномалии могут преобразовываться в новые сигнатуры, обогащая базу знаний системы [6].

Однако важно учитывать, что сами алгоритмы МО, используемые в СОВ, уязвимы к специально сконструированным состязательным атакам (рис. 2) [7].

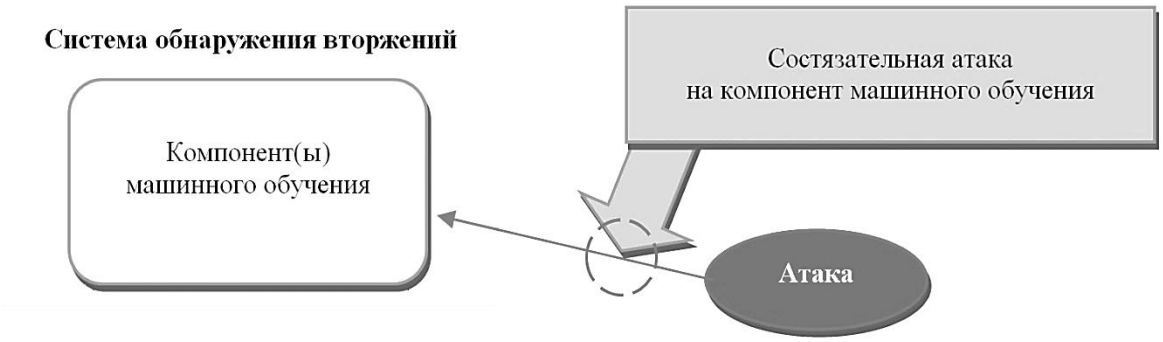

Рис. 2. Состязательная атака на компоненты машинного обучения

Fig. 2. An adversarial attack on machine learning components

Интеграция алгоритмов МО в СОВ важных инфраструктур предъявляет повышенные требования к их устойчивости к современным киберугрозам. Особую актуальность эта проблема приобретает в контексте защиты объектов, отказ которых может вызвать каскадные последствия – от масштабных экономических потерь до угроз национальной безопасности [8].

Структура статьи отражает систематический характер исследования:

– аналитический обзор современных МО-решений в области обнаружения вторжений, введение системы метрик эффективности атак, описание датасетов моделирования;

– детализация механизмов атак на МО-компоненты;

– представление экспериментальных результатов;

– выводы и перспективные направления развития.

Основной фокус исследования сосредоточен на анализе уязвимостей МО-компонентов СОВ через призму состязательных атак. В работе рассмотрен комплексный подход к моделированию таких угроз с последующей оценкой их воздействия.

Набор данных и метрики оценки СОВ с МО

Современные научные публикации убедительно демонстрируют эффективность алгоритмов МО в задачах выявления киберугроз [9]. Подобные результаты подчеркивают актуальность дальнейших исследований, направленных на повышение устойчивости МО-моделей в СОВ к состязательным атакам. Многочисленные работы посвящены разработке СОВ с использованием МО-компонентов обнаружения [9–12]. В частности, исследователи [10] успешно применили метод OC-SVM (One-Class Support Vector Machines, одноклассовые машины с опорными векторами) для классификации и фильтрации сетевого трафика. В другом исследовании [11] нейросетевые алгоритмы продемонстрировали свою эффективность при детектировании ботнет-активности. Интересные результаты были получены при обучении модели RF (Random Forest, случайный лес) на датасете CICIDS – алгоритм показал исключительную точность на тестовых выборках.

В работе [12] авторы провели сравнительный анализ нескольких МО-методов для кибербезопасности, включая DBN (Deep Belief Network, глубокая сеть доверия) и MLP (Multi-layer Perceptron, многослойный персептрон). Важным этапом исследования стало определение информативных признаков и снижение размерности данных, что в итоге позволило достичь высокой точности классификации атак.

Сравнительные характеристики рассмотренных СОВ с МО-компонентами, протестированных на данных CICIDS, приведены в табл. 1.

Таблица 1

Table 1

Сравнение СОВ с компонентами машинного обучения

Comparing IDS with machine learning components

Модель	Метрики, %			СОВ
	F	***Precision***	***Recall***	
OC-SVM / RF	99	99	98	Multi-Stage IDS with Machine Learning Component (Multi-Stage IDS) [10]
Random Forest	97	98	96	Machine Learning-Based Intrusion Detection System (ML-Based IDS) [11]
DBN	94	89	99	Intrusion detection system based on deep learning neural networks (IDS based DBN) [12]

Несмотря на детальную проработку алгоритмов МО в СОВ и тщательный анализ показателей их эффективности, ключевой вопрос остается без ответа: как именно атаки на подрыв устойчивости (Adversarial attacks) искажают метрики детектирования угроз?

Существующие исследования [9–12], хотя и содержат исчерпывающие данные о производительности моделей (таких как OC-SVM, RF или DBN) в стандартных условиях, практически не затрагивают проблему их уязвимости к преднамеренным искажениям входных данных. Между тем в реальных условиях злоумышленники могут целенаправленно манипулировать сетевым трафиком или другими входными параметрами, чтобы обмануть систему защиты. Например, даже высокие показатели точности моделей на тестовых выборках (что демонстрируют работы [11, 12]) не гарантируют их устойчивости к adversarial примерам – специально модифицированным данным, которые вызывают ошибочные предсказания. Это создает серьезный пробел в современных исследованиях, поскольку без учета подобных атак оценка надежности системы может быть существенно завышена.

Таким образом, несмотря на значительный прогресс в применении МО для кибербезопасности критически важными направлениями дальнейших исследований остаются изучение влияния состязательных атак на метрики обнаружения (такие как полнота, точность, *F-мера*) и разработка методов повышения устойчивости моделей к подобным воздействиям.

Для анализа устойчивости МО-компонентов в СОВ к состязательным атакам целесообразно применять те же метрики качества, что и при стандартной оценке эффективности (см. табл. 1). Такой подход обеспечивает сопоставимость результатов и позволяет количественно оценить деградацию системы защиты после воздействия злоумышленника [13].

Ключевым показателем остается точность (*Precision*) – способность системы корректно идентифицировать атаки среди всех срабатываний:

$$Precision = \frac{TP}{TP + FP},$$

где *TP* (True Positives) – верно обнаруженные атаки; *FP* (False Positives) – ошибочные тревоги.

Не менее важна полнота (*Recall*), характеризующая долю выявленных угроз от общего числа атак в данных:

$$Recall = \frac{TP}{TP + FN},$$

где *FN* (False Negatives) – пропущенные угрозы.

Для комплексной оценки применяется *F-мера* – гармоническое среднее *Precision* и *Recall*:

$$F = 2\frac{Precision \cdot Recall}{Precision + Recall}.$$

Эти метрики [14], традиционно используемые для оценки МО-моделей в СОВ, приобретают особое значение при анализе устойчивости: значительное падение *Precision* после атаки указывает на рост ложных срабатываний, а снижение *Recall* свидетельствует о повышении уровня пропуска реальных угроз. Мониторинг этих показателей позволяет не только выявить уязвимости алгоритмов, но и обосновать необходимость разработки специализированных механизмов защиты для развертывания СОВ в важных инфраструктурах.

Датасет CICIDS представляет собой современный и детализированный набор данных, содержащий реалистичные сценарии кибератак, воспроизведенные в условиях, приближенных к реальной сетевой инфраструктуре. Особую ценность этому ресурсу придает тщательная разметка сетевых потоков с указанием временных меток, IP-адресов (Internet Protocol, интернет-протокол), узлов, номеров портов, используемых протоколов и конкретных типов атак.

Для генерации естественного фонового трафика в датасете применена инновационная система *b*-профилей, моделирующая поведенческие паттерны 25 условных пользователей. В основу моделирования легли такие распространенные протоколы, как:

– HTTP (Hypertext Transfer Protocol, протокол передачи гипертекста);

– HTTPS (Hyper Text Transfer Protocol Secure, защищенный протокол передачи гипертекста);

– FTP (File Transfer Protocol, протокол передачи файлов);

– SSH (Secure Shell, безопасная оболочка).

Период сбора информации охватывал рабочую неделю – с 9 : 00 3 июля (понедельник) до 17 : 00 7 июля (пятница) 2017 г. Первый день эксперимента содержал исключительно легитимную активность, в последующие дни внедрялись различные типы атак.

Тестовая среда включала полномасштабную сетевую инфраструктуру с модемом, межсетевым экраном, коммутатором, маршрутизатором и хостами под управлением ОС Ubuntu [15]. Такой комплексный подход к формированию датасета обеспечивает высокую репрезентативность данных для исследований в области информационной безопасности.

Атаки на МО-компоненты СОВ

Атаки на МО-компоненты СОВ представляют собой особый класс угроз, специфически нацеленных на уязвимости алгоритмов искусственного интеллекта в защитных системах [16]. В отличие от традиционных кибератак, эти воздействия эксплуатируют фундаментальные особенности работы нейросетевых моделей и статистических классификаторов, что требует принципиально иных подходов к их анализу и противодействию.

Ключевые особенности таких атак:

– они используют сложность архитектуры нейронных сетей и вероятностный характер их решений для целенаправленного искажения результатов работы;

– стандартные средства киберзащиты (антивирусы, системы сигнатурного анализа) оказываются неэффективными против подобных воздействий;

– атаки могут быть незаметны при традиционном тестировании, но критически влияют на работоспособность системы в реальных условиях [17].

Теоретические предпосылки уязвимостей МО-классификаторов:

– гипотеза нелинейности – сложные нейросетевые модели создают в пространстве признаков области с непредсказуемым поведением, которые могут быть использованы злоумышленниками;

– гипотеза переобучения – излишняя адаптация к тренировочным данным делает модель чувствительной к минимальным отклонениям во входных параметрах [18].

Этот особый характер угроз требует разработки специализированных методов защиты, учитывающих как специфику МО, так и требования к безопасности сложных ИТ-инфраструктур.

Дадим определение состязательной атаки применительно к проблеме классификации данных. Рассмотрим входные данные X и соответствующие им метки классов Y, где модель реализует функцию отображения $f : x \to y$. В этом контексте состязательная атака представляет собой злонамеренное воздействие, при котором противник стремится модифицировать входные данные или метки так, чтобы нарушить корректность работы классификатора f [19]. Такие атаки могут реализовываться разными методами:

– внесение зашумленных возмущений в исходные признаки;

– манипуляция ключевыми параметрами данных;

– преднамеренная подмена меток в обучающей выборке.

Главная задача атакующего – снизить доверие к модели, спровоцировав ошибочные предсказания или системные сбои в ее работе. Математически задача классификации выражается уравнением

$$f : X \to Y,$$

где X – исходный набор данных; Y – конечное множество меток классов.

Отображение f считается уязвимым к состязательным атакам, если существует преобразование A, такое, что для произвольного $x \in X$ можно построить модифицированный пример $\tilde{x} = A(x)$, при котором $f(\tilde{x}) \neq y$, хотя исходно $f(x) = y$. Атака задается оператором:

$$A : R^d \to R^d,$$

где $x \in R^d$ принадлежит классу y, а $\tilde{x} = A(x)$ приводит к ошибочной классификации $f(\tilde{x}) \neq y$. Аддитивные атаки (возмущения) – наиболее распространенный тип состязательных атак, заключающийся в добавлении малого шума $\eta \in R^d$ к исходным данным:

$$\eta \in R^d \text{ к } x, \text{ так что } \tilde{x} = x + \eta \text{ и } f(\tilde{x}) = y_t,$$

причем $y_t \neq y$. Такие атаки сохраняют структуру входного пространства и обладают интерпретируемостью.

Для систем классификации с метками, которые можно записать множеством

$$\{y1, y2, y3, ..., yk\},$$

и множеством решающих функций

$$\{g1(\cdot), g2(\cdot), ..., gk(\cdot)\}.$$

Задача состоит в нарушении работы классификатора так, чтобы f сопоставлял x класс y_t, для этого необходимо, чтобы $g_t(\tilde{x}) \geq g_i(\tilde{x})$, для всех $i \neq t$, тогда значение $g_t(x)$ должно быть больше, чем значение любой другой $g(\cdot)$. Цель атаки – обеспечить выполнение условия

$$g_t(\tilde{x}) \geq \max_{i \neq t} \{g_i(\tilde{x})\} \Leftrightarrow \max_{i \neq t} \{g_i(\tilde{x})\} - g_t(\tilde{x}) \leq 0.$$

Искажение η должно быть как можно меньше, а $\tilde{x}$ должен быть как можно ближе к x. Оптимизационные формулировки атак:

1. Атакующий стремится найти минимальное искажение $\tilde{x}$, нарушающее классификацию (минимизация возмущения):

$$\min_x |x - \tilde{x}| \text{ при } \max_{i \neq t} \{g_i(\tilde{x})\} - g_t(\tilde{x}) \leq 0,$$

где $|.|$ – любая функция расстояния.

2. Альтернативный подход – максимизация ошибки классификации при ограничении на величину возмущения:

$$\min_x \max_{i \neq t} \{g_i(\tilde{x})\} - g_t(\tilde{x}) \text{ при } |x - \tilde{x}| \leq \lambda,$$

где $\lambda > 0$ задает верхнюю границу.

3. Компромисс между двумя критериями достигается введением параметра α (атака с регуляризацией):

$$\min_x |x - \tilde{x}| + \alpha \, (\max_{i \neq t} \{g_i(\tilde{x})\} - g_t(\tilde{x})),$$

где $\alpha > 0$ задает верхнюю границу.

Помимо модификации входных данных существуют атаки на этап обучения, включающие изменение меток существующих данных, добавление вредоносных размеченных примеров [20].

В контексте нарушения работы классификаторов можно выделить 4 ключевые стратегии, различающиеся по уровню доступа атакующего к данным и алгоритмам:

1. Модификация меток. Злоумышленник изменяет только метки классов в размеченной обучающей выборке, не затрагивая сами данные. Это позволяет незаметно искажать решающие правила модели, т. к. изменения касаются исключительно целевых переменных Y.

2. Внедрение данных. При отсутствии прямого доступа к исходному обучающему набору атакующий добавляет в него новые вредоносные примеры. Такие «зараженные» данные, поданные на этапе обучения, могут смещать границы классификации в нужном направлении.

3. Модификация данных. Атакующий получает полный контроль над обучающей выборкой (но не над алгоритмом) и напрямую изменяет признаки X. Это более агрессивный метод, чем внедрение, т. к. он предполагает целенаправленное искажение существующих данных.

4. Логические искажения. Наиболее мощный тип атак, при котором злоумышленник вмешивается в работу самого алгоритма обучения (например, модифицируя функцию потерь или процедуру оптимизации).

Эти стратегии объединяются под термином «отравляющие атаки» (Poisoning attacks), поскольку их цель – не локальное искажение входных данных, а глобальное изменение решающей поверхности классификатора. В отличие от состязательных атак, которые модифицируют отдельные примеры, отравление смещает всю гиперплоскость принятия решений [21].

Во время состязательной атаки злоумышленник изменяет исходные данные путем добавления или вычитания маленькой величины (эпсилона), умноженной на знак градиента. Это делается для максимизации функции потерь (в случае атаки на увеличение предсказанного класса) или для минимизации функции потерь (в случае атаки на уменьшение предсказанного класса).

В процессе атаки целевая модель подвергается воздействию состязательных примеров – специально модифицированных данных, которые, несмотря на кажущуюся схожесть с исходными образцами, приводят к ошибочной классификации. Одним из наиболее известных подходов к генерации таких примеров выступает FGSM (Fast Gradient Sign Method, быстрый метод градиентного знака) – мощный инструмент для обхода защитных механизмов систем МО, включая глубокие нейронные сети. Суть FGSM заключается в формировании адверсарных образцов за счет анализа градиента функции потерь (будь то MSE – среднеквадратичная ошибка или CCE – категориальная кросс-энтропия) относительно входных данных. Направленное искажение вносится в соответствии с полученными градиентными значениями, что позволяет искусственно усилить ошибку предсказания [22]. Визуально модифицированные данные практически неотличимы от настоящих, однако модель выдает совершенно некорректные результаты. Математически метод выражается формулой

$$adv_x = x + \varepsilon \cdot \text{sign}(\nabla_x J(\theta, x, y)),$$

где adv_x – состязательный образец; x – исходный входной образец; ε – маленькое значение, на которое умножаются знаковые градиенты для обеспечения незаметности возмущений, но достаточной величины для обмана нейронной сети; J – функция потерь; θ – модель нейронной сети; y – метка класса исходного образца [23].

Эксперименты по моделированию состязательной атаки FGSM на МО-компоненты СОВ

В данном исследовании рассматриваются методы компрометации СОВ, оснащенных алгоритмами МО. Рассматриваемая ML-Based IDS разработана на Python с привлечением таких широко известных фреймворков, как scikit-learn, TensorFlow и Keras. Tensorflow – основной инструмент для работы с глубинным обучением. Для построения последовательных нейросетевых архитектур применяется библиотека: Keras.

Для детектирования вредоносных URL (Uniform Resource Locator, стандартизированный способ указания местоположения веб-ресурсов) применяется нейросетевая модель, код которой содержит функции: build_model формирует структуру нейронной сети. На входной слой (main_input) подается последовательность целочисленных значений, соответствующих индексам слов в словаре. Далее следует слой Embedding, выполняющий преобразование числовых индексов в векторные представления заданной размерности (emb_dim). Более подробно устройство СОВ описано в [11].

Для тестирования уязвимостей системы использовалась атака FGSM, реализованная на Python [24]. Воздействие данного метода на машинные компоненты СОВ детально проанализировано, а его результаты визуализированы в табл. 2 и на рис. 3.

Таблица 2

Table 2

FGSM-атака на ML-Based IDS, %

FGSM attack on ML-Based IDS, %

ε	*Recall*	*Precision*	*F*
0	98	99	98
0,05	81	98	88
0,10	76	95	84
0,15	71	92	80
0,20	66	89	75
0,25	61	86	71
0,30	56	84	67

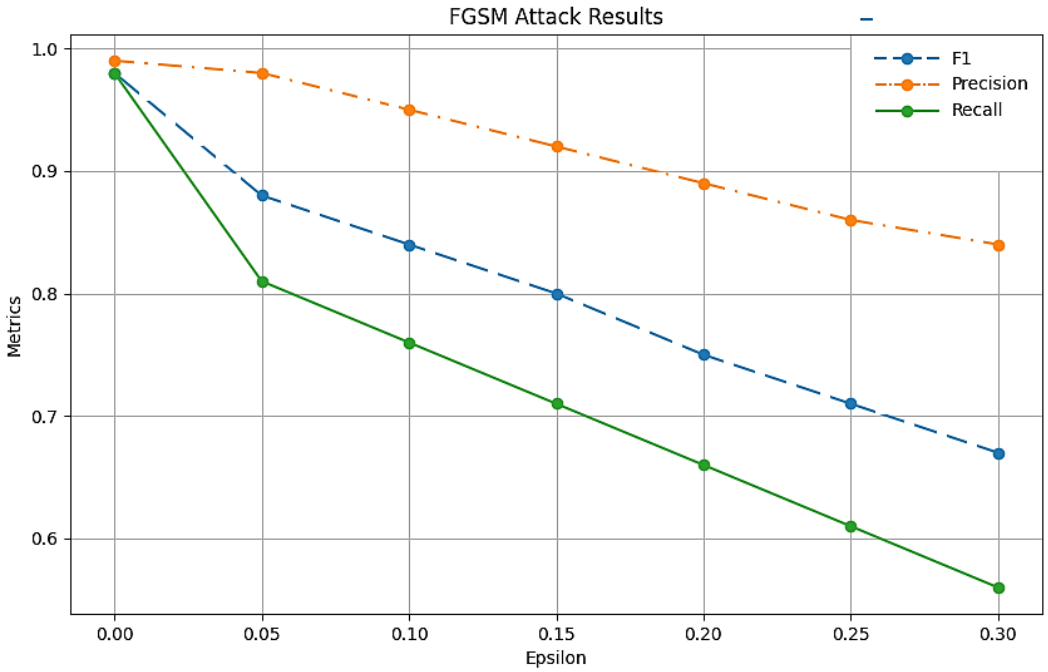

Рис. 3. Результаты атаки на ML-Based IDS

Fig. 3. The results of the attack on ML-Based IDS

Даже минимальная корректировка параметра ε провоцирует существенные колебания в показателях эффективности. Так, при ε = 0,30 наблюдается значение *F-меры* = 67 %, при этом отмечается выраженный дисбаланс между *Precision* = 84 % и *Recall* = 56 %, что явно свидетельствует о дестабилизации системы под воздействием атакующего вмешательства.

В архитектуре Multi-Stage IDS применяется многоуровневая схема выявления аномалий, основанная на иерархическом анализе угроз. Для реализации системы задействован Python с использованием таких инструментов МО, как OneClassSVM (метод одноклассовой классификации), RandomForest (ансамблевый алгоритм на основе решающих деревьев), Classifier (базовый классификатор), предобученные пайплайны и модели, сериализованные посредством библиотеки Pickle [10].

Экспериментальные данные, демонстрирующие последствия атаки на МО-компонент Multi-Stage IDS, детализированы в табл. 3 и визуализированы на рис. 4.

Таблица 3

Table 3

FGSM-атака на Multi-Stage IDS, %

FGSM attack on MultiStage IDS, %

ε	*Recall*	*Precision*	*F*
0	98	99	98
0,05	92	92	92
0,10	88	84	86
0,15	76	72	74
0,20	72	64	68
0,25	68	56	61
0,30	64	52	57

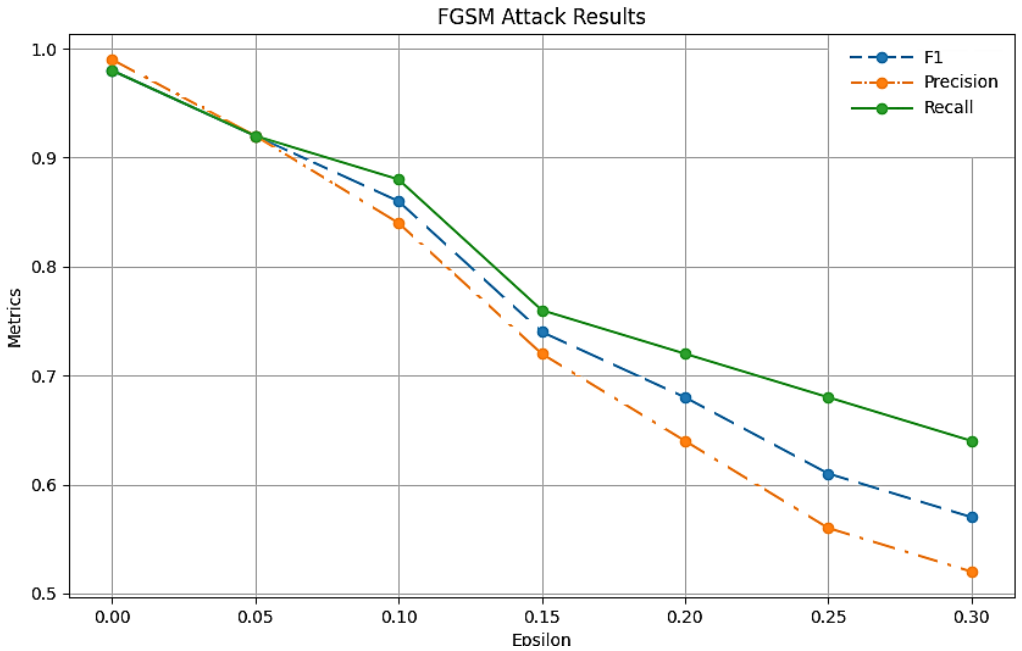

Рис. 4. Результаты атаки на Multi-Stage IDS

Fig. 4. The results of the attack on MultiStage IDS

В данном случае воздействие атаки на МО-компонент проявляется более выраженно по сравнению с предыдущим экспериментом. При значении параметра ε = 0,30 наблюдаются следующие показатели: *F-мера* = 57 %, *Precision* = 52 %, *Recall* = 64 %. Хотя система сохраняет работоспособность, отмечается заметное снижение точности классификации при умеренном нарушении баланса между метриками.

Архитектура СОВ на основе DBN использует принцип построения глубоких нейросетевых моделей через каскад RBM (Restricted Boltzmann Machines, ограниченные машины Больцмана). В реализации задействованы следующие инструменты: библиотека Torch обеспечивает тензорные вычисления, а модуль NN позволяет конструировать слои нейронной сети [12].

Инициализация сети включает последовательное создание RBM-слоев в итеративном режиме. Процедура обучения DBN осуществляется через Python-метод fit, применяемый к каждому RBM-блоку, с последовательным преобразованием входных данных для следующих уровней иерархии. Критерием остановки служит величина MSE (среднеквадратичной ошибки).

Экспериментальные данные, демонстрирующие эффект атаки на IDS based DBN, систематизированы в табл. 4 и на рис. 5.

Таблица 4

Table 4

FGSM-атака на IDS based DBN, %

FGSM attack on IOS based DBN, %

ε	*Recall*	*Precision*	*F*
0	99	88	93
0,05	96	84	90
0,10	92	74	82
0,15	88	72	79
0,20	82	64	72
0,25	80	62	70
0,30	72	60	65

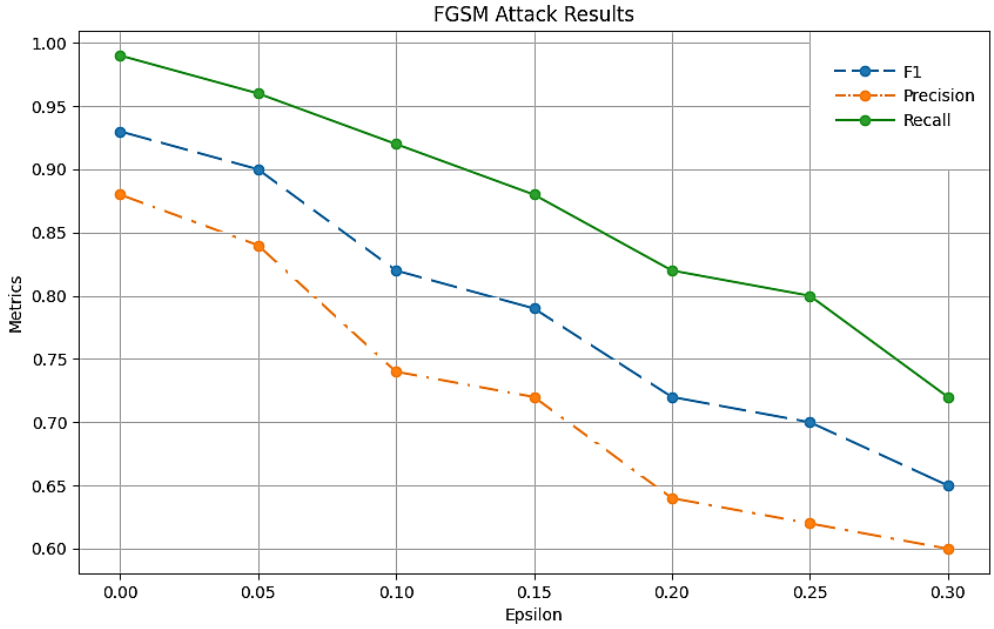

Рис. 5. Результаты атаки на IDS based DBN

Fig. 5. The results of the attack on IDS based DBN

В ходе моделирования рассматриваемой атаки при значении параметра ε = 0,30 были зафиксированы следующие показатели эффективности: *F-мера* достигла 65 %, точность *Precision* составила 60 %, а полнота *Recall* – 72 %. Следует отметить, что реализованная атака способствует балансировке системы, уменьшая разрыв между полнотой и точностью. При детальном анализе выявлено, что воздействие на показатель *Recall* оказывается менее выраженным по сравнению с влиянием на *Precision* – точность системы демонстрирует более существенные изменения под воздействием атаки.

Заключение

В данном исследовании осуществлен комплексный разбор проблемы обеспечения безопасности компонентов машинного обучения (МО) систем обнаружения вторжений (СОВ). Были смоделированы состязательные атаки методом быстрого градиентного знака (FGSM), а их воздействие проанализировано с использованием стандартных оценочных показателей: *Precision* (точность), *Recall* (полнота) и *F-мера*.

Проведенный эксперимент позволил оценить уязвимость к состязательным атакам современных СОВ. Выявлено, что подобные атаки существенно искажают ключевые метрики детектирования, приводя:

– к падению *Recall* (способности выявлять реальные угрозы);

– снижению *Precision* (росту ложных срабатываний);

– нарушению баланса работы всей системы.

Ни одна из протестированных систем не продемонстрировала достаточной резистентности к adversarial-атакам. Полученные результаты актуализируют необходимость разработки специализированных защитных механизмов для МО-компонентов СОВ. Перспективным направлением представляется создание новых моделей и алгоритмов, устойчивых к целенаправленным искажениям входных данных.

Список источников

1. Musa U., Chhabra M., Ali A., Kaur M. Intrusion Detection System using Machine Learning Techniques: A Review // Proceedings of the 2020 International Conference on Smart Electronics and Communication (ICOSEC) (Trichy, India, 10–12 September 2020). P. 149–155. DOI: 10.1109/ICOSEC49089.2020.9215333.

2. Elhanashi A., Gasmi K., Begni A., Dini P., Zheng Q., Saponara S. Machine Learning Techniques for Anomaly-Based Detection System on CSE-CIC-IDS2018 Dataset // Applications in Electronics Pervading Industry, Environment and Society: APPLEPIES 2022. Berlin/Heidelberg, Germany: Springer, 2023. P. 131–140. DOI: 10.1007/978-3-031-30333-3_17.

3. 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.

4. Buschlinger L., Rieke R., Sarda S., Krauß C. Decision Tree-Based Rule Derivation for Intrusion Detection in Safety-Critical Automotive Systems // Proceedings of the 30th Euromicro International Conference on Parallel, Distributed and Network-Based Processing (PDP). March 2022. P. 246–254. DOI: 10.1109/PDP55904.2022.00046.

5. Jing D., Chen H. B. SVM Based Network Intrusion Detection for the UNSW-NB15 Dataset // Proceedings of the 2019 IEEE 13th International Conference on ASIC (ASICON) (Chongqing, China, 29 October–1 November 2019). P. 1–4. DOI: 10.1109/ASICON47005.2019.8983598.

6. Hassine K., Erbad A., Hamila R. Important Complexity Reduction of Random Forest in Multi-Classification Problem // Proceedings of the 2019 15th International Wireless Communications & Mobile Computing Conference (IWCMC) (Tangier, Morocco, 24–28 June 2019). P. 226–231. DOI: 10.1109/IWCMC.2019.8766544.

7. Ageev S., Kopchak Y., Kotenko I., Saenko I. Abnormal Traffic Detection in Networks of the Internet of Things Based on Fuzzy Logical Inference // Proceedings of International Conference on Soft Computing and Measurements (SCM 2015). 2015. P. 5–8. DOI: 10.1109/SCM.2015.7190394.

8. Branitskiy A., Kotenko I. Network Attack Detection Based on Combination of Neural, Immune and Neuro-Fuzzy Classifiers // Proceedings of the IEEE 18th International Conference on Computational Science and Engineering (CSE 2015). 2015. P. 152–159. DOI: 10.15217/issn1684-8853.2015.4.69.

9. Kotenko I., Kuleshov A., Ushakov I. Aggregation of Elastic Stack Instruments for Collecting, Storing and Processing of Security Information and Events // Proceedings of the 14th IEEE Conference on Advanced and Trusted Computing (ATC 2017) (San Francisco, USA, 4–8 August 2017). P. 1550–1557. DOI: 10.1109/UIC-ATC.2017.8397627.

10. Verkerken M., D'hooge L., Sudyana D., Lin Y., Wauters T., Volckaert B., De Turck F. Novel Multi-Stage Approach for Hierarchical Intrusion Detection // IEEE Transactions on Network and Service Management. 2023. P. 1–1. DOI: 10.1109/TNSM.2023.3259474.

11. Goryunov M., Matskevich A., Rybolovlev D. Synthesis of a Machine Learning Model for Detecting Computer Attacks Based on the CICIDS2017 Dataset // Trudy ISP RAN/Proc. ISP RAS. 2020. V. 32. Iss. 5. P. 81–94. DOI: 10.15514/ISPRAS-2020-32(5)-7.

12. Belarbi O., Khan A., Carnelli P., Spyridopoulos T. An Intrusion Detection System Based on Deep Belief Networks // Proceedings of the 4th International Conference on Science of Cyber Security (SciSec 2022). Cham: Springer, 2022. P. 377–392. DOI: 10.1007/978-3-031-17551-0_25.

13. Primartha R., Tama B. A. Anomaly Detection Using Random Forest: A Performance Revisited // Proceedings of the 2017 International Conference on Data and Software Engineering (ICoDSE) (Palembang, Indonesia, 1–2 November 2017). P. 1–6. DOI: 10.1109/ICODSE.2017.8285847.

14. Kalaivaani P. T., Krishnamoorthy R., Reddy A. S. D., Chelladurai A. D. D. Adaptive Multimode Decision Tree Classification Model Using Effective System Analysis in IDS for 5G and IoT Security Issues // Secure Communication for 5G and IoT Networks. Springer, 2022. P. 141–158. DOI: 10.1007/978-3-030-79766-9_9.

15. Panigrahi R., Borah S. A Detailed Analysis of CICIDS2017 Dataset for Designing Intrusion Detection Systems // International Journal of Engineering & Technology. 2018. V. 7. N. 3.24. P. 479–482.

16. Solani S., Jadav N. A Novel Approach to Reduce False-Negative Alarm Rate in Network-Based Intrusion Detection System Using Linear Discriminant Analysis //

Inventive Communication and Computational Technologies. Berlin/Heidelberg, Germany: Springer, 2021. P. 911–921. DOI: 10.1007/978-981-15-7345-3_77.

17. Hanif S., Ilyas T., Zeeshan M. Intrusion Detection in IoT Using Artificial Neural Networks on UNSW-15 Dataset // Proceedings of the 2019 IEEE 16th International Conference on Smart Cities: Improving Quality of Life Using ICT IoT and AI (HONET-ICT) (Charlotte, NC, USA, 6–9 October 2019). P. 152–156. DOI: 10.1109/HONET.2019.8908122.

18. Rajagopal S., Hareesha K., Kundapur P. Feature Relevance Analysis and Feature Reduction of UNSW NB-15 Using Neural Networks on MAMLS // Advanced Computing and Intelligent Engineering. Berlin/Heidelberg, Germany: Springer, 2020. P. 321–332. DOI: 10.1007/978-981-15-1483-8_28.

19. Kanimozhi V., Jacob T. Artificial Intelligence Based Network Intrusion Detection with Hyper-Parameter Optimization Tuning on the Realistic Cyber Dataset CSE-CIC-IDS2018 Using Cloud Computing // Proceedings of the 2019 International Conference on Communication and Signal Processing (ICCSP) (Chennai, India, 4–6 April 2019). P. 33–36. DOI: 10.1109/ICCSP.2019.8698055.

20. Ilyushin E., Namiot D., Chizhov I. Attacks on Machine Learning Systems - Common Problems and Methods // International Journal of Open Information Technologies. 2022. V. 10. N. 3. P. 48–54.

21. Devarakonda A., Sharma N., Saha P., Ramya S. Network Intrusion Detection: A Comparative Study of Four Classifiers Using the NSL-KDD and KDD'99 Datasets // Journal of Physics: Conference Series. 2022. V. 2161. P. 012043. DOI: 10.1088/1742-6596/2161/1/012043.

22. Wiyono S., Abidin T. Comparative Study of Machine Learning KNN, SVM, and Decision Tree Algorithm to Predict Student's Performance. International Journal of Research Granthaalayah. 2019. V. 7. N. 1. P. 190–196. DOI: 10.29121/granthaalayah.v7.i1.2019.1009.

23. Moualla S., Khorzom K., Jafar A. Improving the Performance of Machine Learning-Based Network Intrusion Detection Systems on the UNSW-NB15 Dataset // Computational Intelligence and Neuroscience. 2021. V. 2021. P. 352–375. DOI: 10.1155/2021/5557577.

24. Ichetovkin E., Kotenko I. Modeling Attacks on Machine Learning Components of Intrusion Detection Systems // 2024 International SmartIndustryCon (Sochi, Russian Federation, 2024). P. 261–266. DOI: 10.1109/SmartIndustryCon 61328.2024.10515506.

References

1. Musa U., Chhabra M., Ali A., Kaur M. Intrusion Detection System using Machine Learning Techniques: A Review. *Proceedings of the 2020 International Conference on Smart Electronics and Communication (ICOSEC) (Trichy, India, 10–12 September 2020)*. Pp. 149-155. DOI: 10.1109/ICOSEC49089.2020.9215333.

2. Elhanashi A., Gasmi K., Begni A., Dini P., Zheng Q., Saponara S. Machine Learning Techniques for Anomaly-Based Detection System on CSE-CIC-IDS2018 Dataset. *Applications in Electronics Pervading Industry, Environment and Society: APPLEPIES 2022.* Berlin/Heidelberg, Germany: Springer, 2023. Pp. 131-140. DOI: 10.1007/978-3-031-30333-3_17.

3. 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).* Pp. 1850-1855. DOI: 10.1109/EDM61683.2024.10615198.

4. Buschlinger L., Rieke R., Sarda S., Krauß C. Decision Tree-Based Rule Derivation for Intrusion Detection in Safety-Critical Automotive Systems. *Proceedings of the 30th Euromicro International Conference on Parallel, Distributed and Network-Based Processing (PDP).* March 2022. Pp. 246-254. DOI: 10.1109/PDP55904.2022.00046.

5. Jing D., Chen H. B. SVM Based Network Intrusion Detection for the UNSW-NB15 Dataset. *Proceedings of the 2019 IEEE 13th International Conference on ASIC (ASICON) (Chongqing, China, 29 October–1 November 2019)*. Pp. 1-4. DOI: 10.1109/ASICON47005.2019.8983598.

6. Hassine K., Erbad A., Hamila R. Important Complexity Reduction of Random Forest in Multi-Classification Problem. *Proceedings of the 2019 15th International Wireless Communications & Mobile Computing Conference (IWCMC) (Tangier, Morocco, 24–28 June 2019).* Pp. 226-231. DOI: 10.1109/IWCMC.2019.8766544.

7. Ageev S., Kopchak Y., Kotenko I., Saenko I. Abnormal Traffic Detection in Networks of the Internet of Things Based on Fuzzy Logical Inference. *Proceedings of International Conference on Soft Computing and Measurements (SCM 2015).* 2015. Pp. 5-8. DOI: 10.1109/SCM.2015.7190394.

8. Branitskiy A., Kotenko I. Network Attack Detection Based on Combination of Neural, Immune and Neuro-Fuzzy Classifiers. *Proceedings of the IEEE 18th International Conference on Computational Science and Engineering (CSE 2015)*. 2015. Pp. 152-159. DOI: 10.15217/issn1684-8853.2015.4.69.

9. Kotenko I., Kuleshov A., Ushakov I. Aggregation of Elastic Stack Instruments for Collecting, Storing and Processing of Security Information and Events. *Proceedings of the 14th IEEE Conference on Advanced and Trusted Computing (ATC 2017) (San Francisco, USA, 4–8 August 2017).* Pp. 1550-1557. DOI: 10.1109/UIC-ATC.2017.8397627.

10. Verkerken M., D'hooge L., Sudyana D., Lin Y., Wauters T., Volckaert B., De Turck F. Novel Multi-Stage Approach for Hierarchical Intrusion Detection. *IEEE Transactions on Network and Service Management*. 2023. Pp. 1-1. DOI: 10.1109/TNSM.2023.3259474.

11. Goryunov M., Matskevich A., Rybolovlev D. Synthesis of a Machine Learning Model for Detecting Computer Attacks Based on the CICIDS2017 Dataset. *Trudy ISP RAN/Proc. ISP RAS*, 2020, vol. 32, iss. 5, pp. 81-94. DOI: 10.15514/ISPRAS-2020-32(5)-7.

12. Belarbi O., Khan A., Carnelli P., Spyridopoulos T. An Intrusion Detection System Based on Deep Belief Networks. *Proceedings of the 4th International Conference on Science of Cyber Security (SciSec 2022).* Cham, Springer, 2022. Pp. 377-392. DOI: 10.1007/978-3-031-17551-0_25.

13. Primartha R., Tama B. A. Anomaly Detection Using Random Forest: A Performance Revisited. *Proceedings of the 2017 International Conference on Data and Software Engineering (ICoDSE) (Palembang, Indonesia, 1–2 November 2017)*. Pp. 1-6. DOI: 10.1109/ICODSE.2017.8285847.

14. Kalaivaani P. T., Krishnamoorthy R., Reddy A. S. D., Chelladurai A. D. D. Adaptive Multimode Decision Tree Classification Model Using Effective System Analysis in IDS for 5G and IoT Security Issues. *Secure Communication for 5G and IoT Networks.* Springer, 2022. Pp. 141-158. DOI: 10.1007/978-3-030-79766-9_9.

15. Panigrahi R., Borah S. A Detailed Analysis of CICIDS2017 Dataset for Designing Intrusion Detection Systems. *International Journal of Engineering & Technology*, 2018, vol. 7, no. 3.24, pp. 479-482.

16. Solani S., Jadav N. A Novel Approach to Reduce False-Negative Alarm Rate in Network-Based Intrusion Detection System Using Linear Discriminant Analysis. *Inventive Communication and Computational Technologies.* Berlin/Heidelberg, Germany, Springer, 2021. Pp. 911-921. DOI: 10.1007/978-981-15-7345-3_77.

17. Hanif S., Ilyas T., Zeeshan M. Intrusion Detection in IoT Using Artificial Neural Networks on UNSW-15 Dataset. *Proceedings of the 2019 IEEE 16th International Conference on Smart Cities: Improving Quality of Life Using ICT IoT and AI (HONET-ICT) (Charlotte, NC, USA, 6-9 October 2019).* Pp. 152-156. DOI: 10.1109/HONET.2019.8908122.

18. Rajagopal S., Hareesha K., Kundapur P. Feature Relevance Analysis and Feature Reduction of UNSW NB-15 Using Neural Networks on MAMLS. *Advanced Computing and Intelligent Engineering*. Berlin/Heidelberg, Germany, Springer, 2020. Pp. 321-332. DOI: 10.1007/978-981-15-1483-8_28.

19. Kanimozhi V., Jacob T. Artificial Intelligence Based Network Intrusion Detection with Hyper-Parameter Optimization Tuning on the Realistic Cyber Dataset CSE-CIC-IDS2018 Using Cloud Computing. *Proceedings of the 2019 International Conference on Communication and Signal Processing (ICCSP) (Chennai, India, 4–6 April 2019).* Pp. 33-36. DOI: 10.1109/ICCSP.2019.8698055.

20. Ilyushin E., Namiot D., Chizhov I. Attacks on Machine Learning Systems - Common Problems and Methods. *International Journal of Open Information Technologies*, 2022, vol. 10, no. 3, pp. 48-54.

21. Devarakonda A., Sharma N., Saha P., Ramya S. Network Intrusion Detection: A Comparative Study of Four Classifiers Using the NSL-KDD and KDD'99 Datasets. *Journal of Physics: Conference Series*, 2022, vol. 2161, p. 012043. DOI: 10.1088/1742-6596/2161/1/012043.

22. Wiyono S., Abidin T. Comparative Study of Machine Learning KNN, SVM, and Decision Tree Algorithm to Predict Student's Performance. *International Journal of Research Granthaalayah*, 2019, vol. 7, no. 1, pp. 190-196. DOI: 10.29121/granthaalayah.v7.i1.2019.1009.

23. Moualla S., Khorzom K., Jafar A. Improving the Performance of Machine Learning-Based Network Intrusion Detection Systems on the UNSW-NB15 Dataset. *Computational Intelligence and Neuroscience*, 2021, vol. 2021, pp. 352-375. DOI: 10.1155/2021/5557577.

24. Ichetovkin E., Kotenko I. Modeling Attacks on Machine Learning Components of Intrusion Detection Systems. *2024 International SmartIndustryCon (Sochi, Russian Federation, 2024).* Pp. 261-266. DOI: 10.1109/SmartIndustryCon 61328.2024.10515506.

Статья поступила в редакцию 21.03.2025; одобрена после рецензирования 18.04.2025; принята к публикации 30.04.2025
The article was submitted 21.03.2025; approved after reviewing 18.04.2025; accepted for publication 30.04.2025

Информация об авторе / Information about the author

Егор Андреевич Ичетовкин – аспирант лаборатории проблем компьютерной безопасности; Санкт-Петербургский Федеральный исследовательский центр Российской академии наук; ichetovkin.e@iias.spb.su

Egor A. Ichetovkin – Postgraduate Student of the Computer Security Laboratories; St. Petersburg Federal Research Center of the Russian Academy of Sciences; ichetovkin.e@iias.spb.su