DOI 10.37882/2223-2966.2025.04-2.14

МОДЕЛИ И АЛГОРИТМЫ АТАК ОТРАВЛЕНИЕМ НА КОМПОНЕНТЫ МАШИННОГО ОБУЧЕНИЯ СИСТЕМ ОБНАРУЖЕНИЯ ВТОРЖЕНИЙ

MODELS AND ALGORITHMS FOR POISONING ATTACKS ON MACHINE LEARNING COMPONENTS OF INTRUSION DETECTION SYSTEMS

Ichetovkin E.

Summary. In modern conditions, intrusion detection systems integrating machine learning methods are actively used to detect cyber threats in computer networks. Such solutions demonstrate high efficiency in detecting anomalies, but their reliability can be significantly reduced in the case of poisoning attacks aimed at compromising training. This study provides a detailed analysis of models and algorithms of such attacks, and assesses their impact on the operation of defense mechanisms. The methodological basis of the work includes modeling data poisoning attacks with subsequent assessment of system performance using classical metrics: accuracy, recall, and F-measure. The key novelty of the study lies in its integrated approach. Various machine learning algorithms used to detect anomalies are considered, including: single-class support vector machines, random forest, and deep machine learning. For the first time, comparative testing of several systems under the influence of targeted poisoning attacks was carried out, which made it possible to identify their critical vulnerabilities. The results of the experimental analysis confirmed that the studied systems are susceptible to poisoning attacks, which leads to a significant decrease in their detection ability. In particular, a 15–30 % drop in F-score was observed depending on the attack type and the model used. The findings highlight the need to develop more robust learning methods that are resistant to adversarial influences, as well as to modernize existing intrusion detection systems with intelligent classifiers. The work contributes to the development of defense mechanisms by offering not only threat analysis, but also a direction for further research in the field of improving the resilience of machine learning algorithms for intrusion detection systems under poisoning attacks.

Keywords: cybersecurity, intrusion detection systems, poisoning attacks, machine learning component, models and algorithms.

Ичетовкин Егор Андреевич
Аспирант, Лаборатория проблем компьютерной безопасности, Санкт-Петербургский Федеральный исследовательский центр Российской академии наук
ichetovkin.e@iias.spb.su

Аннотация. В современных условиях для выявления киберугроз в компьютерных сетях активно применяются системы обнаружения вторжений, интегрирующие методы машинного обучения. Подобные решения демонстрируют высокую эффективность при детектировании аномалий, однако их надежность может быть существенно снижена в случае реализации атак отравления, направленных на компрометацию обучения. В данном исследовании проводится детальный анализ моделей и алгоритмов подобных атак, а также оценивается их влияние на работу защитных механизмов. *Методологическая основа* работы включает моделирование атак отравления данных с последующей оценкой производительности систем с использованием классических метрик: точности, полноты и F-меры. *Ключевая новизна* исследования заключается в комплексном подходе. Рассматриваются различные алгоритмы машинного обучения, применяемые для обнаружения аномалий, включая: одноклассовые машины с опорными векторами, случайный лес и глубокое машинное обучение. *Впервые* проведено сравнительное тестирование нескольких систем под воздействием целенаправленных атак отравлением, что позволило выявить их критические уязвимости. *Результаты* экспериментального анализа подтвердили, что исследуемые системы подвержены влиянию атак отравления, что приводит к значительному снижению их детектирующей способности. В частности, наблюдалось падение F-меры на 15–30 % в зависимости от типа атаки и используемой модели. Полученные данные подчеркивают необходимость разработки более устойчивых методов обучения, устойчивых к состязательным воздействиям, а также модернизации существующих систем обнаружения вторжений с интеллектуальными классификаторами. *Работа вносит вклад* в развитие защитных механизмов, предлагая не только анализ угроз, но и направление для дальнейших исследований в области повышения устойчивости алгоритмов машинного обучения систем обнаружения вторжений в условиях атак отравлением.

Ключевые слова: кибербезопасность, системы обнаружения вторжений, атаки отравлением, компонент машинного обучения, модели и алгоритмы.

Введение

В современных реалиях обеспечение информационной безопасности немыслимо без систем обнаружения вторжений (СОВ). Рост числа и сложности кибератак на важную инфраструктуру, включая объекты государственного значения, требует разработки более эффективных методов защиты [1].

Применяемые для защиты ответственной инфраструктуры СОВ принято делить на три категории:

- Сигнатурные — используют заранее заданные правила для выявления известных угроз, демонстрируя высокую точность при обнаружении атак с известными шаблонами.
- Эвристические — опираются на анализ поведенческих признаков (эвристик) сетевой активности.

Последние исследования подтверждают, что методы машинного обучения (МО) значительно повышают эффективность таких систем.

- Гибридные — комбинируют оба подхода, что позволяет сочетать оперативность сигнатурного анализа с широким охватом эвристических методов [2].

Особую опасность для компонентов машинного обучения гибридных и эвристических СОВ представляют состязательные атаки. Поскольку они направлены на манипуляцию с классификаторами детектирования аномалий [3]. В отличие от традиционных киберугроз, эти атаки эксплуатируют специфику нейросетевых моделей — их сложную структуру и статистическую природу. Стандартные механизмы защиты зачастую оказываются бесполезными против таких атак [4]. Существует несколько объяснений уязвимости моделей машинного обучения систем обнаружения вторжений:

- Нелинейность архитектуры — приводит к формированию «слепых зон», где модель некорректно интерпретирует входные данные.
- Переобучение — даже незначительные отклонения от обучающей выборки могут вызвать ошибочные предсказания.

Атаки на компоненты МО СОВ делятся на основные группы:

1. Атаки уклонением (evasion attacks) — злоумышленник модифицирует входные данные (например, добавляет шум или искажает признаки), чтобы обмануть классификатор.
2. Атаки отравлением (poisoning attacks) — направлены не на входные данные, а на саму модель, изменяя её параметры или границы принятия решений.
3. Скрытые угрозы (backdoors/trojans) — особая разновидность атак отравлением, при которой модель тайно активирует вредоносную логику при определенных условиях.

В работе анализируются методы и алгоритмы poisoning-атак (отравления), представляющих серьезную угрозу для систем машинного обучения, особенно в условиях ограниченности обучающих выборок [5].

Основная задача исследования — моделирование и анализ воздействия атак отравления на МО-компоненты систем обнаружения вторжений. Особое внимание уделяется исследованию динамики ключевых метрик качества классификации: *precision* (точности), *recall* (полноты) и *F*-меры при целенаправленном искажении обучающих данных.

Научная новизна исследования заключается в систематизации критериев оценки уязвимости классификаторов СОВ, разработке математической модели атак отравления и детальной спецификации экспериментальной методики. Приведены табличные данные и графики, демонстрирующие влияние искаженных данных на метрики МО-моделей.

Экспериментальная часть включает количественную оценку деградации качества классификации по трем ключевым показателям в условиях различных сценариев атак.

Структура статьи построена следующим образом: в Разделе I обсуждаются основные критерии оценки эффективности классификаторов. Раздел II посвящен сравнительному анализу различных алгоритмов компонентов машинного обучения. Описание датасетов и их характеристик приведено в Разделе III. Математические модели атак формализованы в Разделах IV–V. Методология эксперимента изложена в Разделе VI, тогда как Раздел VII содержит интерпретацию полученных результатов и выводы.

I. Критерии обнаружения

Эффективность классификаторов принято оценивать с помощью специальных метрик, которые показывают, насколько точно система отличает реальные угрозы от ошибочных тревог. Проще говоря, эти показатели помогают понять, сколько атак действительно обнаружила система и сколько из них оказались ложными срабатываниями [6].

Одним из ключевых показателей является точность (*Precision*) — доля корректно выявленных атак среди всех срабатываний системы. Формула расчета выглядит следующим образом:

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

где TP (True Positives) — верно определенные атаки, а FP (False Positives) — ложные срабатывания.

Еще одна важная метрика — полнота (*Recall*), показывающая, какая часть реальных атак была обнаружена:

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

где FN (False Negatives) означает пропущенные угрозы.

Для комплексной оценки используют F-меру — гармоническое среднее между точностью и полнотой:

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

Данный показатель учитывает как ложные срабатывания (FP), так и пропущенные атаки (FN), что делает его

особенно полезным при анализе сбалансированности модели [7].

II. Системы обнаружения вторжений на базе МО

Современные научные работы [8–10] подтверждают эффективность алгоритмов машинного обучения в задачах детектирования киберугроз. Подобные выводы актуализируют потребность в дальнейшем изучении уязвимостей самих классификаторов к целенаправленным атакам.

В контексте классификации сетевого трафика исследователи [11] добились значимых результатов, применив метод OC-SVM (One-Class Support Vector Machines, одноклассовые машины с опорными векторами). Другой подход продемонстрирован в работе [12], где нейросетевые технологии использовались для выявления ботнет-активности — здесь модель RF (Random Forest, случайный лес), обученная на датасете CICIDS, показала исключительную детектирующую способность.

Не менее интересны сравнительные исследования алгоритмов машинного обучения в области кибербезопасности. Так, авторы [13] провели комплексный анализ эффективности различных архитектур, включая DBN (Deep Belief Network, глубинные сети доверия) и MLP (Multi-layer Perceptron, многослойный перцептрон). Ключевыми этапами стали предварительный отбор и снижение размерности пространства информативных признаков.

Сводные характеристики систем обнаружения вторжений с компонентами машинного обучения систематизированы в таблице 1.

Таблица 1.

Сравнение СОВ с компонентами машинного обучения

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]

Рассматриваемые параметры характеризуют интеграцию алгоритмов машинного обучения в СОВ и включают соответствующие метрики эффективности детектирования угроз. Вместе с тем, авторы не затрагивают критически важный аспект — уязвимость классификаторов к атакам отравлением. Для восполнения этого пробела требуется разработка специализированной системы оценивания устойчивости моделей к подобным воздействиям, как обосновано в исследовании [14].

III. Набор данных для моделирования атак

В современных исследованиях по информационной безопасности широко применяется датасет CICIDS [15], включающий как модели реальных кибератак, так и примеры нормальной сетевой активности. Этот комплексный набор данных предоставляет детализированную информацию о сетевых потоках, включая временные метки, IP-адреса (Internet Protocol, интернет-протокол) взаимодействующих узлов, номера портов, используемые протоколы и классификацию атакующих методик.

Особенностью CICIDS является применение системы b-профилей, которая анализирует типичные паттерны поведения пользователей и генерирует соответствующий фоновый трафик. В рамках датасета смоделирована деятельность 25 абстрактных пользователей, взаимодействующих через распространенные протоколы: 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, защищенная оболочка), а также использующих электронную почту.

Тестовая среда включала типовые сетевые компоненты: модем, межсетевой экран, коммутатор, маршрутизатор и хосты под управлением ОС Ubuntu [16].

Данные охватывают непрерывную неделю мониторинга, что позволяет анализировать динамику сетевой активности с привязкой ко времени и дням недели. В текущей работе CICIDS используется для оценки систем из таблицы 1 по критериям (1–3) в двух состояниях — до и после атаки. Такой подход дает возможность измерить воздействие кибератак на производительность систем обнаружения вторжений с компонентами машинного обучения [17].

IV. Модели атак отравлением

В отличие от атак уклонением, атаки отравлением манипулируют обучающими данными. Эти атаки можно классифицировать по четырём ключевым типам, различающимся по возможностям злоумышленника и моменту вмешательства. При этом атакующий может столкнуться с ограничениями: допустимое число модификаций, штрафы за изменения, допустимые типы преобразований данных [18].

Одним из распространённых методов является атака модификации меток (Label modification attacks), при которой изменяются только метки данных. Эта атака применима к произвольным точкам данных, но обычно имеет ограничения, например, на максимальное число изменяемых меток. Чаще всего она используется против двоичных классификаторов [19].

Другой тип — атака вставки яда (Poison insertion attacks), когда злоумышленник добавляет вредоносные векторы признаков. В отличие от предыдущего метода, эта атака работает в условиях обучения без учителя, где можно манипулировать только признаками, но не метками [20].

Более гибкой является атака модификации данных (Data modification attacks), позволяющая изменять как признаки, так и метки для произвольного подмножества обучающих данных. Это делает её особенно опасной, поскольку атакующий может адаптировать изменения под конкретную модель [21].

Наконец, атака «кипящей лягушки» (Boiling frog attacks) основана на постепенном отравлении модели в процессе её итеративного переобучения. В отличие от агрессивных вмешательств, здесь злоумышленник вносит малозаметные изменения на каждом этапе, что в долгосрочной перспективе серьёзно искажает работу модели. Такая атака особенно эффективна в сценариях обучения без учителя [22].

В типичном сценарии атак с отравлением данных злонамеренные акторы первоначально работают с нетронутым обучающим набором D_0, который впоследствии модифицируется в отравленную версию D. На этом изменённом наборе затем происходит обучение целевой модели машинного обучения [23].

Подобно атакам на этапе инференса, на этапе обучения злоумышленники могут преследовать два типа целей:

- Целевые атаки, направленные на манипуляцию предсказаниями модели для конкретных выборок S.
- Атаки на устойчивость, целью которых является общее снижение точности модели.

Ключевой задачей атакующего становится балансирование между эффективностью атаки и незаметностью внесённых изменений. Математически это выражается через функцию риска:

$$R_A(D, S), \tag{4}$$

которая зависит от параметров модели W, обученных на модифицированных данных D.

Стоимость модификации набора данных описывается уравнением:

$$c(D_0, D). \tag{5}$$

Оптимизационная задача злоумышленника часто формулируется как:

$$\min_D R_A(D, S) + \lambda c(D_0, D), \tag{6}$$

где параметр λ регулирует баланс между успешностью атаки и масштабом вмешательства. Альтернативно, можно использовать ограниченную по ресурсам постановку:

$$\min_D R_A(D, S),$$
$$c(D_0, D) \le C, \tag{7}$$

где C — максимально допустимая стоимость модификаций.

В ряде случаев вместо минимизации риска удобнее максимизировать функцию успешности атаки, которая связана с R_A обратной зависимостью, что позволяет напрямую оценить эффективность атаки:

$$U_A(D, S) = -R_A(D, S), \tag{8}$$

где U — успешность атаки, которую киберпреступник пытаться максимизировать.

Таким образом, выбор между этими подходами зависит от конкретных условий атаки и доступных злоумышленнику ресурсов. [24].

V. Алгоритмы атак отравлением

Атаки, нацеленные на системы онлайн-детектирования, представляют особую угрозу, поскольку нарушают работу эвристических методов анализа и затрудняют выявление редких или ранее неизвестных угроз. Речь идет о так называемой BFA (Boiling frog attacks, атака «кипящая лягушка»), которая основана на предположении, что алгоритм обнаружения аномалий регулярно обновляет свою модель по мере поступления новых данных. Злоумышленник постепенно накапливает информацию и внедряет изменения между последовательными циклами переобучения модели.

Ключевой элемент этой атаки — целевой вектор признаков X_T, который злоумышленник стремится внедрить в систему так, чтобы в дальнейшем он был ошибочно классифицирован как легитимный. По сути, это медленная и незаметная манипуляция, напоминающая принцип «лягушки, не замечающей постепенного нагрева воды» [25].

На определенном шаге обучения T атакующая стратегия достигает успеха, если выполняется условие:

$$\|X_T - \mu_T \geq r\|, \quad (9)$$

где T — номер итерации, μ_T — текущее положение центроида, а r — заданный порог детектирования аномалий.

В отличие от случайных возмущений, эта атака носит целенаправленный характер: злоумышленник стремится подобрать такое X_T, чтобы расстояние до центроида равнялось пороговому значению:

$$\|X_T - \mu_T\| = r. \quad (10)$$

Подобная тактика, основанная на постепенном «отравлении» обучающей выборки, является уникальной для онлайн-детекторов, использующих центроиды. Ее ключевое преимущество — скрытое внедрение вредоносных образцов, которые формально остаются в рамках нормы, но систематически смещают границы принятия решений [26].

Механизм атаки можно представить следующим образом. На каждой итерации, когда у злоумышленника появляется возможность модификации данных, он добавляет объект X_T расположенный строго на границе допустимой области. Вычисляется направление смещения:

$$a = \frac{X_T - \mu_T}{\|X_T - \mu_T\|}, \quad (11)$$

единичный вектор, который задает ориентацию атаки относительно центроида, но при этом остается в пределах нормальной области.

Тогда сам атакующий экземпляр принимает вид:

$$X_A = \mu_T + ra, \quad (12)$$

Такой подход гарантирует, что внедряемые данные не вызовут срабатывания детектора, но приведут к постепенной деградации его точности. В результате, когда произойдет реальная атака, система может классифицировать ее как нормальное поведение, что создает серьезные угрозы безопасности [27].

VI. Моделирование атак отравлением

Ранее был представлен сравнительный анализ систем обнаружения вторжений с компонентами машинного обучения, характеристика которых систематизированными в таблице 1. В рамках исследования изучалась устойчивость данных систем к атакам отравление. Работа классификаторов СОВ оценивалась по трем ключевым метрикам: полнота, точность, F-мера (1–3).

Для моделирования атак был выбран алгоритм BFA (атака «кипящая лягушка»). Для успешной реализации таких атак на МО-компоненты СОВ требуется выполнение двух критических условий.

Во-первых, злоумышленник должен постоянно мониторить оптимальные решения X_A.

Во-вторых, если в процессе онлайн-атаки нарушается равенство (10), необходимо заменять X_T на μ_T для его восстановления.

Суть BFA-атаки заключается в поиске минимального возмущения, способного изменить результат классификации входных данных [28].

В качестве примера рассмотрен Python-алгоритм, который может быть адаптирован для манипуляции процессом распознавания сетевых атак. Алгоритм маркирует все соединения как атаки, если они происходят в заданном временном окне после известного вредоносного события. Код реализации ранее был представлен автором в исследование [6]. Применяя в заранее определенные временные интервалы X_T, атака модифицирует исходно «чистый» трафик μ_T, что в ходе обучения модели приводит к смещению центроида согласно (12).

Это, в свою очередь, вызывает снижение эффективности СОВ, что проявляется в ухудшении значений метрик (1-3). Здесь параметр X_T отражает долю вредоносного трафика, искусственно внедренного в нормальный трафик μ_T. При превышении порогового значения r система перестает адекватно реагировать на аномальную активность, теряя способность к детектированию угроз.

VII. Эксперимент

Проанализируем уязвимость ML-based IDS. Исходный код системы, как и всех представленных в работе, разработан на высокоуровневом языке программирования Python и представлен в открытом доступе в сети Интернет.

Система реализована с применением следующих библиотек машинного обучения: scikit-learn для классических алгоритмов, tensorflow и keras — для глубокого обучения [11]. В частности, из пакета keras задействованы:

- модуль sequential, позволяющий строить слои нейросети;
- компонент model, обеспечивающий гибкость архитектуры.

Экспериментальное исследование проводилось с использованием BFA (атака «кипящей лягушки») — метода, который манипулирует обучающими данными.

Результаты тестирования с варьируемыми параметрами X_T демонстрируют уязвимость представленной системы обнаружения вторжений (таблица 2). Визуальная интерпретация динамики атаки представлена на рис. 1, где четко прослеживаются критические точки деградации детектора.

Таблица 2.

BFA атака на ML-based IDS

X_T	μ_T	*F*,%	*Recall*	*Precision*
0.00	1.00	97	96	98
0.06	0.94	95	94	96
0.08	0.92	92	90	95
0.12	0.88	92	88	96
0.16	0.84	88	82	95
0.20	0.80	85	78	94
0.24	0.76	73	62	90
0.28	0.72	64	54	80
0.32	0.68	58	44	85
0.36	0.64	48	34	80
0.40	0.60	35	24	66
0.44	0.56	23	14	62
0.48	0.52	4	02	38
0.52	0.48	4	02	34

Анализ экспериментальных данных показывает, что даже при незначительном уровне загрязнения трафика (X_T =0.2, что соответствует 20% вредоносных пакетов) наблюдается выраженная деградация ключевых метрик качества классификации.

Интересно отметить, что показатели *Recall* (полнота) демонстрируют более резкий спад по сравнению с *Precision* (точностью), что свидетельствует о растущем количестве ложноположительных срабатываний при увеличении доли аномального трафика.

В рамках работы была исследована многоуровневая система обнаружения вторжений Multi-stage IDS. СОВ основанна на принципах иерархического анализа. Реализация выполнена на Python с активным использованием таких библиотек машинного обучения, как scikit-learn (включая модули One-Class SVM, Random Forest и классификаторы), а также инструментов сериализации данных (pickle) для интеграции компонентов [12].

Влияние на работоспособность МО-компонентов системы обнаружения вторжений в рамках проведенной BFA-атака видно из таблицы 3 и рис. 2.

На рис. 2 наблюдается сбалансированность системы, что подтверждается устойчивыми показателями *F*-меры. Тем не менее, при достижении уровня вредоносного трафика X_T =0.16 (16%) отмечается снижение точности, которое при X_T =0.27 приводит к практически полному выравниванию её абсолютного значения с полнотой.

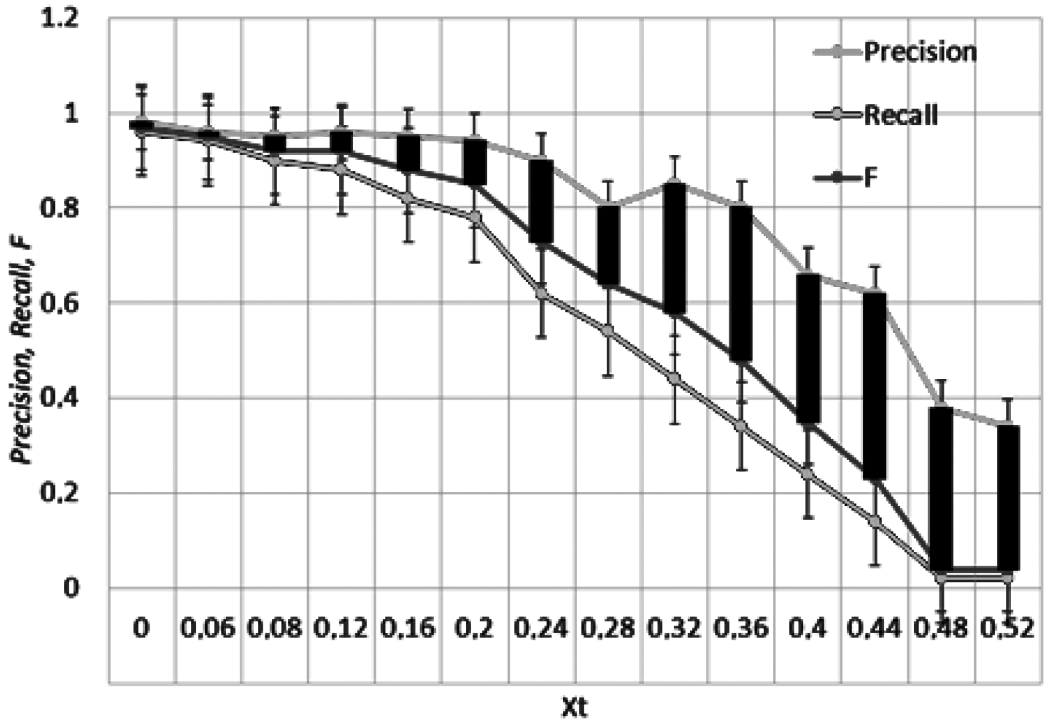

Рис. 1. Результат BFA атаки на ML-based IDS

Таблица 3.

BFA атака на Multi-stage IDS

X_T	μ_T	*F*,%	*Recall*	*Precision*
0.00	1.00	98	98	99
0.06	0.94	97	96	98
0.08	0.92	93	90	96
0.12	0.88	89	87	92
0.16	0.84	86	83	90
0.20	0.80	82	79	86
0.24	0.76	77	72	82
0.28	0.72	64	64	65
0.32	0.68	56	57	56
0.36	0.64	48	50	47
0.40	0.60	40	42	38
0.44	0.56	31	35	29
0.48	0.52	23	28	20
0.52	0.48	14	20	11

Для следующего эксперимента была выбрана IDS based DBN это СОВ на основе DBN (Deep Belief Network, глубокая доверительная сеть) — архитектуры машинного обучения, построенной на комбинации нейронных сетей глубокого обучения и RBM (Restricted Boltzmann Machines, ограниченных машин Больцмана), выполняющих роль скрытых слоёв. Реализация системы выполнена на Python с задействованием библиотек Torch и NN. Первая обеспечивает операции с тензорами и алгоритмами глубокого обучения, а модуль NN в PyTorch применяется для конструирования нейросетевых моделей.

Инициализация RBM для скрытых слоёв осуществляется последовательно: в цикле генерируются экземпляры RBM, где первый слой связывается с входными данными, а последующие — друг с другом. Обучение DBN проводится через метод fit для каждой RBM с параллельным обновлением входных данных для следующих слоёв сети, что в итоге возвращает значение среднеквадратичной ошибки [13].

Данные по атакам на МО-компоненты системы обнаружения вторжений приведены в таблице 4 и проиллюстрированы на рис. 3.

Данная атака демонстрирует нетипичное поведение, характеризующееся множественными пересечениями кривых *Precision* и *Recall* в точках X_T =0.24, X_T =0.32 и X_T =0.48. Такая динамика, вероятно, обусловлена спецификой функционирования DBN. Несмотря на общую сбалансированность системы и устойчивость метрики *F*, наблюдается критическое ухудшение всех ключевых показателей (*Precision*, *Recall*, *F*) при превышении порога X_T=0.2 (эквивалентно 20 % доли вредоносного трафика в условно легитимном потоке).

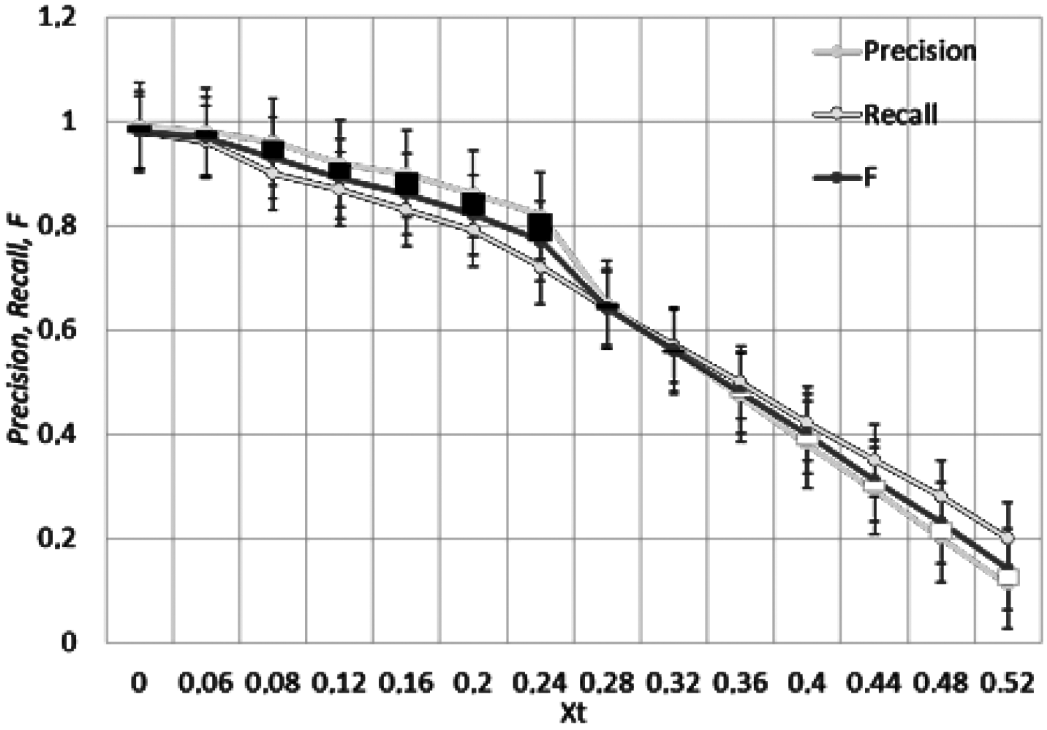

Рис. 2. Результат BFA атаки на Multi-stage IDS

Таблица 4.

BFA атака на IDS based DBN

X_T	μ_T	F,%	Recall	Precision
0.00	1.00	94	99	89
0.06	0.94	93	98	88
0.08	0.92	89	92	86
0.12	0.88	77	83	72
0.16	0.84	77	80	74
0.20	0.80	70	74	66
0.24	0.76	68	69	68
0.28	0.72	66	63	70
0.32	0.68	58	58	59
0.36	0.64	44	44	44
0.40	0.60	33	41	28
0.44	0.56	20	34	14
0.48	0.52	12	12	12
0.52	0.48	2	1	10

Проведённый анализ систем обнаружения вторжений выявил их критическую зависимость от устойчивости алгоритмов машинного обучения к целенаправленным искажениям входных данных. Как показывают эксперименты, все рассмотренные решения демонстрируют susceptibility (восприимчивость) к poisoning attacks (атакам на заражение обучающей выборки), что ставит под сомнение их надёжность при защите critical infrastructure (ответственных объектов информационной инфраструктуры).

Для практического использования таких систем в реальных условиях необходимо обязательное внедрение специализированных защитных механизмов — своеобразного «иммунного ответа» на попытки компрометации данных. Только при наличии таких countermeasures (противодействий) можно гарантировать, что технологии машинного обучения смогут эффективно выполнять свою защитную функцию, выступая в роли «цифровых дозорных» на стратегически важных участках сети.

Примечательно, что эта уязвимость носит фундаментальный характер — она присуща даже наиболее продвинутым hybrid IDS (гибридным системам обнаружения вторжений), сочетающим сигнатурные и поведенческие методы анализа. Это напоминает ситуацию, когда самая совершенная охранная сигнализация остаётся бесполезной, если злоумышленник получает доступ к её настройкам.

Заключение

В данной работе представлен всесторонний анализ уязвимостей современных алгоритмов машинного обучения, применяемых в системах обнаружения вторжений. Для этого представлены модели и алгоритмы атак отравления. Проведенные эксперименты убедительно

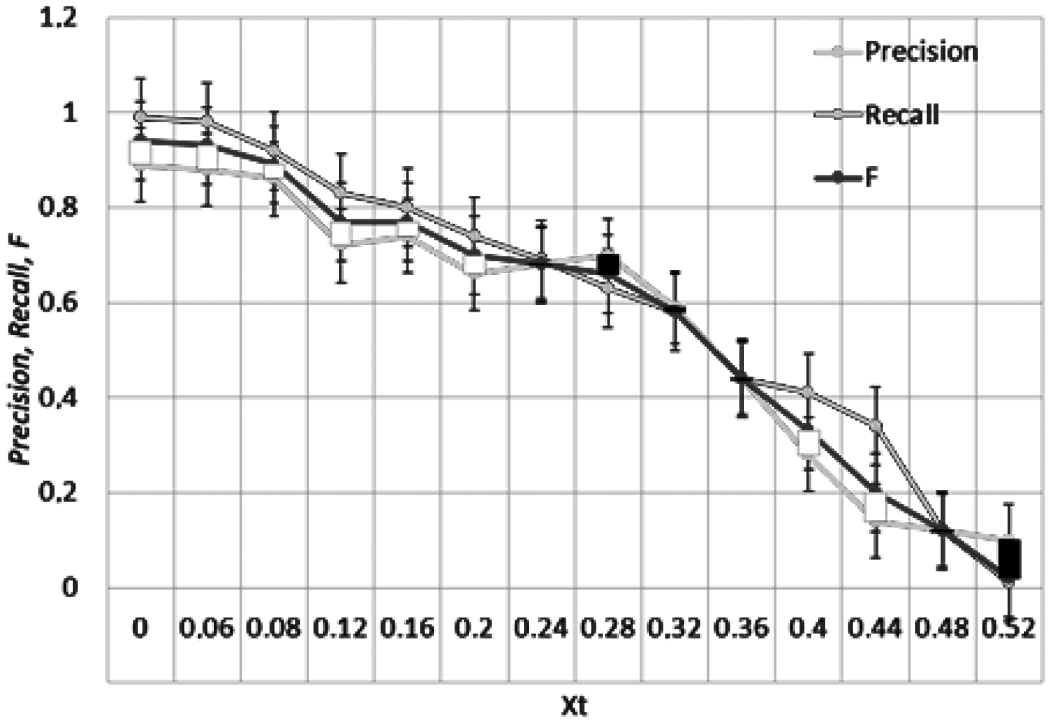

Рис. 3. Результат BFA атаки на IDS based DBN

демонстрируют парадоксальную ситуацию: хотя методы искусственного интеллекта действительно эффективно выявляют кибератаки, сами алгоритмы классификации оказываются крайне восприимчивыми к целенаправленным манипуляциям с обучающими данными.

Особую опасность представляет так называемый «эффект лягушки в кипятке» (boiling frog attack), при котором злоумышленник постепенно вносит незаметные изменения в обучающую выборку. Исследование показало, что подобные атаки приводят к значительной деградации ключевых метрик качества работы классификаторов — точности (*precision*), полноты (*recall*) и *F*-меры, которые являются критически важными для оценки эффективности систем информационной безопасности.

В качестве потенциального решения проблемы предлагается использовать архитектуры на основе автоэнкодеров (autoencoders) — специальных нейросетевых моделей, способных выявлять аномалии в данных. Такие системы могут служить дополнительным защитным механизмом, фильтруя потенциально опасные образцы перед их подачей на вход основного классификатора.

Перспективы дальнейших исследований видятся в следующих направлениях:

- Разработка новых алгоритмических подходов к защите компонентов машинного обучения СОВ.
- Исследование других типов состязательных атак, в частности атак уклонения (evasion attacks).
- Создание комплексных методов оценки устойчивости моделей к различным видам адверсарных воздействий.

Полученные результаты подчеркивают необходимость создания специализированных механизмов защиты компонентов машинного обучения систем обнаружения вторжений, работающих в условиях потенциальных киберугроз.

ЛИТЕРАТУРА

1. 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_12.
2. 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/ASICON47005.2019.8983564.
3. 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/ATC.2017.8167682.
4. 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/SmartIndustryCon61328.2024.10515506.
5. 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.2019.8908059.
6. 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.
7. Alshehri A., Alghamdi A., Alqahtani A. A Machine Learning-Based Intrusion Detection for Detecting Internet of Things Network Attacks // Journal of King Saud University — Computer and Information Sciences. 2022. DOI: 10.1016/j.jksuci.2022.05.005.
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: 10.1109/CSE.2015.28.
9. Kotenko I., Chechulin A., Komashinsky D. Categorisation of Web Pages for Protection Against Inappropriate Content in the Internet // International Journal of Internet Protocol Technology. 2017. Vol. 10, No. 1. P. 61–71. DOI: 10.1504/IJIPT.2017.083381.
10. Branitskiy A., Kotenko I. Hybridization of Computational Intelligence Methods for Attack Detection in Computer Networks // Journal of Computational Science. 2017. Vol. 23. P. 145–156. DOI: 10.1016/j.jocs.2017.10.003.
11. 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.2023.3318705.
12. 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, Issue 5. P. 81–94. DOI: 10.15514/ISPRAS-2020-32(5)-6.
13. 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: 10.1007/978-3-031-17189-5_26.
14. Zhang C., Costa-Pérez X., Patras P. Adversarial Attacks Against Deep Learning-Based Network Intrusion Detection Systems and Defense Mechanisms // IEEE/ACM Transactions on Networking. 2022. Vol. 30, No. 3. P. 1294–1311. DOI: 10.1109/TNET.2022.3159581.
15. Sauka K., Shin G.-Y., Kim D.-W., Han M.-M. Adversarial Robust and Explainable Network Intrusion Detection Systems Based on Deep Learning // Applied Sciences. 2022. Vol. 12, No. 13. P. 6451. DOI: 10.3390/app12136451.
16. 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.

17. Pierazzi F., Pendlebury F., Cortellazzi J., Cavallaro L. Intriguing Properties of Adversarial ML Attacks in the Problem Space // Proceedings of the IEEE Symposium on Security and Privacy (SP). 2020. P. 1332–1349. DOI: 10.1109/SP40000.2020.00057.
18. Abdelaty M., Scott-Hayward S., Doriguzzi-Corin R., Siracusa D. GaDoT: GAN-Based Adversarial Training for Robust DDoS Attack Detection // Proceedings of the 9th IEEE Conference on Communications and Network Security (CNS). 2021. P. 1–9. DOI: 10.1109/CNS53000.2021.9705029.
19. Tramèr F., Kurakin A., Papernot N., Goodfellow I.J., Boneh D., McDaniel P.D. Ensemble Adversarial Training: Attacks and Defenses // Proceedings of the 6th International Conference on Learning Representations (ICLR). 2018. P. 1–22. DOI: 10.48550/arXiv.1705.07204.
20. Pawlicki M., Choraś M., Kozik R. Defending Network Intrusion Detection Systems Against Adversarial Evasion Attacks // Future Generation Computer Systems. 2020. Vol. 110. P. 148–154. DOI: 10.1016/j.future.2020.04.012.
21. Patel N., Krishnamurthy P., Garg S., Khorrami F. Overriding Autonomous Driving Systems Using Adaptive Adversarial Billboards // IEEE Transactions on Intelligent Transportation Systems. 2022. Vol. 23, No. 8. P. 11386–11396. DOI: 10.1109/TITS.2021.3129785.
22. Vaccari I., Carlevaro A., Narteni S., Cambiaso E., Mongelli M. Explainable and Reliable Against Adversarial Machine Learning in Data Analytics // IEEE Access. 2022. Vol. 10. P. 83949–83970. DOI: 10.1109/ACCESS.2022.3197500.
23. Moosavi-Dezfooli S.-M., Fawzi A., Frossard P. DeepFool: A Simple and Accurate Method to Fool Deep Neural Networks // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016. P. 2574–2582. DOI: 10.1109/CVPR.2016.282.
24. Costantino G., Matteucci I. Reversing Kia Motors Head Unit to Discover and Exploit Software Vulnerabilities // Journal of Computer Virology and Hacking Techniques. 2022. Vol. 19. P. 33–49. DOI: 10.1007/s11416-022-00441-2.
25. Cheng Q., Zhou S., Shen Y., Kong D., Wu C. Packet-Level Adversarial Network Traffic Crafting Using Sequence Generative Adversarial Networks. 2021. P. 47–94. DOI: 10.48550/arXiv.2106.12345.
26. 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.
27. Mozaffari S., Al-Jarrahand O.Y., Dianati M., Jennings P., Mouzakitis A. Deep Learning-Based Vehicle Behaviour Prediction for Autonomous Driving Applications: A Review // IEEE Transactions on Intelligent Transportation Systems. 2022. Vol. 23, No. 1. P. 33–47. DOI: 10.1109/TITS.2021.3117259.
28. Hsu T.-M., Wang C.-H., Hsu H.-C., Chiang C.-H., Chen Y.-R. Motion Planning of Autonomous Vehicle for Illegal Parked Vehicles at No Parking Area // Proceedings of the IEEE International Automatic Control Conference. 2021. P. 1–5. DOI: 10.1109/IACC51493.2021.9609488.

Журнал «Современная наука: актуальные проблемы теории и практики»