Оригинальная статья / Original article

УДК 004.93

https://doi.org/10.21869/2223-1560-2025-29-4-125-139

Применение искусственного интеллекта в задачах обнаружения деструктивных воздействий на информационные и технические системы

Д.Е. Селиверстов [1] ✉, К.Д. Русаков [1]

[1] Институт проблем управления им. В. А. Трапезникова Российской академии наук
ул. Профсоюзная, д. 65, г. Москва 117997, Российская Федерация

✉ e-mail: Seliverstov_dmitriyy@rambler.ru

Резюме

Целью работы *являлось обоснование эффективности применения и сравнение методов искусственного интеллекта (машинного обучения и глубокого обучения) для своевременного обнаружения деструктивных воздействий на информационные и технические системы.*

Методы. *Выполнен анализ современных научных источников, включая обзоры и стандарты по кибербезопасности, а также проведен эксперимент на открытом наборе данных сетевых атак (UNSW-NB15) с использованием алгоритмов машинного обучения (Random Forest) и глубокой нейронной сети. Оценка проводилась по метрикам точности, полноты обнаружения, F1 и др.*

Результаты. *Методы ML/DL демонстрируют существенно более высокую точность обнаружения воздействий по сравнению с традиционными сигнатурными средствами: на датасете UNSW-NB15 достигнута точность ~96% при использовании нейронной сети против ~70% у сигнатурного подхода. Показано, что глубокое обучение позволяет выявлять ранее неизвестные атаки (в т.ч. сложные многовекторные) за счет распознавания скрытых аномалий, а ансамблевые и федеративные подходы повышают надежность и скорость обнаружения.*

Заключение. *Интеграция методов ИИ в системы мониторинга безопасности значительно повышает эффективность защиты информационных и технических систем за счет проактивного выявления кибератак с минимальными ложными срабатываниями. Экспериментальные результаты подтверждают практическую применимость выбранных методов для защиты сетевой инфраструктуры (энергетика, связь, промышленный IoT), однако требуют дальнейшего развития в части обеспечения устойчивости к воздействиям и соблюдения принципов надежности ИИ.*

Ключевые слова: *искусственный интеллект; машинное обучение; глубокое обучение; цифровой двойник; федеративное обучение; обнаружение атак; выявление аномалий; ситуационная осведомленность; автома-тическое реагирование; кибербезопасность; эргатические системы; киберустойчивость.*

Конфликт интересов: *Авторы декларируют отсутствие явных и потенциальных конфликтов интересов, связанных с публикацией настоящей статьи.*

Для цитирования: Селиверстов Д.Е., Русаков К.Д. Применение искусственного интеллекта в задачах обнаружения деструктивных воздействий на информационные и технические системы // Известия Юго-Западного государственного университета. 2025; 29(4): 125-139. https://doi.org/10.21869/-2223-1560-2025-29-4-125-139.

Поступила в редакцию 03.09.2025 *Подписана в печать 14.10.2025* *Опубликована 22.12.2025*

Application of artificial intelligence for detecting information-technical impacts

Dmitry E. Seliverstov [1] ✉, Konstantin D. Rusakov [1]

[1] V. A. Trapeznikov Institute of Control Sciences of the Russian Academy of Sciences
65 Profsoyuznaya str., Moscow 117997, Russian Federation

✉ e-mail: Seliverstov_dmitriyy@rambler.ru

Abstract

Purpose *of the work was to substantiate the effectiveness of applying artificial intelligence techniques (machine learning and deep learning) for the timely detection of destructive information-technical impacts on critical infrastructure objects.*

Methods*. An analysis of scientific sources has been conducted scientific sources, including cybersecurity surveys and standards, and conducted an experiment on a public network attack dataset (UNSW-NB15) using machine learning (Random Forest) and a deep neural network. Evaluation was based on metrics such as accuracy, detection recall, F1-score, etc.*

Results*. ML/DL methods show significantly higher attack detection accuracy compared to traditional signature-based tools: ~96% accuracy was achieved on the UNSW-NB15 dataset using a neural network, versus ~70% for the signature approach. We demonstrate that deep learning enables discovery of previously unknown attacks (including sophisticated multi-vector APTs) by recognizing hidden anomalies, and that ensemble and federated approaches improve detection reliability and speed.* ***Conclusion****. Integrating AI techniques into security monitoring systems considerably increases the protection efficiency of critical systems by proactively identifying cyberattacks with minimal false alarms. The experimental results confirm the practical applicability of the chosen methods for securing network infrastructure (energy, communications, industrial IoT). However, further work is needed to ensure robustness against adversarial attacks and to uphold AI reliability principles.*

Keywords: *artificial intelligence; machine learning; deep learning; attack detection; anomalies; critical infrastructure; cybersecurity.*

Conflict of interest. *The Authors declare the absence of obvious and potential conflicts of interest related to the publication of this article.*

For citation: Seliverstov D. E., Rusakov K. D. Application of Artificial Intelligence for Detecting Information-Technical Impacts. *Izvestiya Yugo-Zapadnogo gosudarstvennogo universiteta = Proceedings of the Southwest State University.* 2025; 29(4): 125-139 (In Russ.). https:// doi.org/10.21869/2223-1560-2025-29-4-125-139.

Received 03.09.2025 *Accepted 14.10.2025* *Published 22.12.2025*

***

Введение

Постоянный рост масштабов цифровизации и усложнение ИТ-инфраструктур приводит к появлению новых угроз информационной безопасности. Деструктивное воздействие на информационные и технические системы понимается как воздействие на информационные ресурсы и технические системы, нарушающее их нормальное функционирование (включая кибератаки, технические сбои,

несанкционированный доступ) [1]. Современные коммуникационные технологии используются злоумышленниками в рамках гибридных атак, способных выводить из строя системы жизнеобеспечения, нарушать работу объектов информационной инфраструктуры и похищать конфиденциальные данные. При этом эргатические системы управления (человеко-машинные комплексы, управляющие процессами в промышленности, транспорте, обороне и др.) особенно уязвимы, поскольку сбои или атаки в таких системах могут иметь катастрофические последствия для государства и общества.

Традиционные средства защиты – от межсетевых экранов до систем обнаружения вторжений (IDS) на основе сигнатур – в последнее время не справляются с динамичным и сложным характером современных атак [2]. Классические IDS/IPS, как правило, централизованы и полагаются на заранее известные шаблоны атак, вследствие чего часто не способны выявлять новые виды атак (например, сложные APT-кампании, инсайдерские угрозы) и генерируют множество ложных срабатываний. Более того, централизованные архитектуры имеют единую точку отказа: компрометация или отказ центрального узла мониторинга выводит из строя всю систему защиты. Это неприемлемо для высоконагруженных систем, где требуется непрерывный мониторинг и реагирование даже в условиях частичных отказов инфраструктуры.

В этой ситуации все большее внимание уделяется использованию методов искусственного интеллекта для усиления средств киберзащиты. Алгоритмы машинного обучения (ML) и глубокого обучения (DL) способны анализировать большие объемы данных о событиях безопасности в режиме реального времени и выявлять скрытые шаблоны, ускользающие от традиционных правил [2]. В отличие от сигнатурных методов, которые опираются на ручное задание признаков угроз, современные методы ML и DL автоматически обучаются на данных и обнаруживают ранее неизвестные атаки по неявным корреляциям признаков. Например, DL-модели способны выявлять сложные многоходовые атаки за счет распознавания малозаметных аномальных характеристик в поведении объектов, чего невозможно добиться ручной экспертизой. Благодаря этому интеграция ИИ позволяет снизить количество ложных тревог и повышает быстродействие обнаружения инцидентов, что особенно важно для ситуационной осведомленности в киберпространстве.

Актуальность исследования. Учитывая рост сложности киберугроз и недостаточность традиционных средств защиты, актуальным является всестороннее изучение и внедрение технологий ИИ для автоматизированного обнаружения деструктивных воздействий на информационные и технические системы. Опыт применения данной технологии свидетельствует о том, что ИИ ста-

новится ключевым элементом современной киберзащиты, позволяя обрабатывать большие данные угроз и реагировать проактивно [3]. В нормативных правовых актах подчеркивается необходимость развития систем противодействия деструктивным воздействиям на информационные и технические системы, включая кибератаки на сложную техническую инфраструктуру[1]. Таким образом, обоснование эффективности применения методов искусственного интеллекта (машинного обучения и глубокого обучения) для своевременного обнаружения таких воздействий и их сравнение имеют важное научное и практическое значение для укрепления национальной и информационной безопасности.

Материалы и методы

Исследование базируется на методах системного анализа и экспериментальной проверки. Выполнен обзор современных публикаций по применению ИИ в кибербезопасности. В том числе рассмотрены обобщающие работы по AI-детекции атак, исследования по аномальному обнаружению и анализ стандартов. Учитывались руководящие документы по безопасности промышленных систем и международные стратегии в области ИИ. Это позволило выработать требования к методам обнаружения атак в сложной технической инфраструктуре (точность, надежность, интерпретируемость результатов, соответствие принципам доверенного ИИ и др.). Кроме того, проведен собственный эксперимент по обнаружению сетевых атак с помощью ML/DL. В качестве данных выбран общедоступный набор UNSW-NB15, имитирующий сетевой трафик с нормальной активностью и несколькими типами атак. Данный датасет содержит около 2,5 млн. сетевых сессий с размеченными атаками 9 различных категорий (DoS, сканирование, эксплойты, бэкдоры, и др.) и нормальный трафик. Для эксперимента использованы: классический алгоритм ML (Random Forest) и модель DL (полносвязная нейронная сеть). Модели обучены классифицировать сетевые сессии по классам (отсутствие либо один из типов атаки). Качество оценивалось на тестовой выборке с помощью метрик точности (accuracy), полноты обнаружения (recall), точности прогноза (precision) и F1-меры. Также рассчитывались показатели для каждого класса атак (категории угроз) и строилась матрица ошибок классификации.

Результаты и их обсуждение

Исторически системы обнаружения вторжений опирались на сигнатурный анализ, при котором известные шаблоны атак сопоставляются с поступающими событиями. Такая схема до сих пор широко распространена в коммерческих IDS/IPS за счет простоты реализации и понятности результатов [4].

[1] Указ Президента РФ от 02.07.2021 № 400 «О Стратегии национальной безопасности Российской Федерации». Официальное опубликование: Официальный интернет-портал правовой информации. URL: https://publication.pravo.gov.ru/Document/View/0001202107030001.

Однако главный недостаток сигнатурного метода – неспособность выявлять новые, ранее не встречавшиеся виды атак. Злоумышленники скрывают вредоносную функциональность и создают эксплойты на незафиксированные уязвимости, сигнатуры на которые отсутствуют, поэтому сигнатурные IDS пропускают такие атаки. Кроме того, рост количества сигнатур ведет к увеличению числа ложных срабатываний и перегружает аналитиков безопасности.

Для преодоления ограничений сигнатурных IDS был развит подход анализа аномалий, при котором модель нормального поведения системы строится на основе статистики, а отклонения от нормы рассматриваются как потенциальные инциденты. Ранние методы аномального обнаружения использовали статистические модели и пороги, но они не учитывали сложных взаимосвязей признаков и часто страдали от высокого числа ложных тревог. Тем не менее, сам переход от поиска известных шаблонов к поиску любых отклонений заложил основу для применения машинного обучения в обнаружении атак [5].

Применение машинного обучения (ML). С середины 2000-х годов в системах обнаружения вторжений начали использовать методы машинного обучения. Классические алгоритмы ML – такие как Decision Tree (деревья решений), Random Forest, SVM (опорные векторы), Naive Bayes, k-NN и др. – обучаются классифицировать сетевые сессии или события на основе множества признаков (особенностей трафика, системных вызовов и пр.). Исследования показывают, что ML-алгоритмы способны достичь высокой точности (>90%) на известных датасетах (например, NSL-KDD) при детектировании известных типов атак. Так, в работе Sowmya & Anita (2023) проведен обзор ряда исследований и сделан вывод, что использование ML повышает точность обнаружения по сравнению с сигнатурным анализом [6, 7]. Однако ML-методы первого поколения имели ограничение: они требовали тщательного ручного конструирования признаков для обучения. Эффективность классификации сильно зависела от качества и полноты выбранных признаков, что затрудняло обнаружение новых атак, не отраженных в признаковых шаблонах. Кроме того, многие ML-модели обучены на несбалансированных данных, вследствие чего им трудно обнаруживать редкие атаки.

Глубокое обучение (DL). Появление глубоких нейронных сетей дало новый импульс развитию IDS. DL-модели (глубокие нейронные сети, рекуррентные сети, автоэнкодеры и др.) автоматически извлекают существенные признаки из сырых данных, устраняя необходимость ручного создания признаков [8]. В кибербезопасности DL находит применение в различных задачах: от обнаружения сетевых атак с помощью рекуррентных нейронных сетей (учитывающих временную динамику трафика) до классификации вредоносных файлов на основе сверточных нейросетей (рас-

познающих паттерны в последовательности байтов). DL продемонстрировал способность выявлять сложные нелинейные зависимости, что особенно эффективно для обнаружения продвинутых угроз. Например, отмечается успех применения DL для противодействия скрытным APT-атакам – нейросети могут распознавать едва заметные последовательности действий атакующего на разных этапах кибер-цепочки [9]. Кроме того, автоэнкодеры применяются для выявления аномалий без меток путем обучения на нормальном поведении: сеть учится восстанавливать «нормальные» данные, и большие ошибки восстановления сигнализируют о нетипичных, возможно, вредоносных образцах. Исследования показывают, что автоэнкодеры успешно обнаруживают атаки на промышленных протоколах, достигая >90% F1-меры в тестовых сценариях [10]. В целом, DL-методы в настоящее время считаются наиболее перспективными для выявления ранее неизвестных атак благодаря их способности обобщать скрытые закономерности. Однако у них есть свои недостатки, например, потребность в больших вычислительных ресурсах и данных для обучения, а также уязвимость к адверсариальным атакам (когда злоумышленник целенаправленно искажает входные данные, чтобы обмануть модель). В сложных технических системах системах эти факторы ограничивают внедрение DL: требуется тщательно тестировать модели на устойчивость и интрепретируемость результатов.

Ансамблевые и гибридные методы. Для повышения надежности часто используют ансамбли алгоритмов и гибридные системы. Ансамблевые классификаторы (например, стекинг или бустинг над разными ML-моделями) позволяют комбинировать сильные стороны отдельных методов и сглаживать их недостатки. Гибридные IDS сочетают в себе одновременно сигнатурные модули для известных атак и аномалические модули на базе ML/DL – тем самым обеспечивая многоуровневую фильтрацию угроз. Например, система сначала отфильтровывает известные атаки сигнатурно, а затем неизвестные отклонения обрабатывает автоэнкодер или кластеризация. Такой подход снижает нагрузку на ML-модель и облегчает интерпретацию. За последние годы отмечается тренд на интеграцию разнородных технологий и переход к многоступенчатым схемам обнаружения [11]. Это особенно актуально в масштабных инфраструктурах, где один подход не обеспечивает полной защиты. Отдельно следует отметить, что повышение надежности интеллектуальных модулей гибридных IDS возможно за счет корректного выбора операций нечеткого вывода и «мягких» арифметических операций, что снижает вычислительные ошибки и стабилизирует принятие решений [12].

Применение ИИ для ситуационной осведомленности. Помимо выявления конкретных атак, ИИ все шире применяется для киберситуационной осведомленности, то есть для формирования це-

лостной картины состояния киберпространства организации в реальном времени. Инструменты на базе ИИ способны агрегировать и анализировать разнородные источники данных: сетевые логи, телеметрию систем, уязвимости, разведданные об угрозах и предоставлять операторам обобщенное представление о текущих инцидентах и рисках. Например, алгоритмы обработки естественного языка (NLP) используются для автоматического анализа текстовых отчетов и новостей о киберугрозах, чтобы выделять релевантную информацию [13]. В итоге, ИИ-инструменты существенно расширяют возможности аналитиков центров мониторинга безопасности: рутинные задачи (поиск индикаторов компрометации, первичная классификация инцидентов) автоматизируются, а персонал может сосредоточиться на принятии решений.

Отечественные разработки. В России активно ведутся исследования по применению ИИ в кибербезопасности. Так, работы под руководством И.В. Котенко посвящены использованию методов ML/DL для защиты IoT и киберфизических систем [14]. В частности, предложены решения по раннему обнаружению кибератак с помощью интеграции статистических методов и фрактального анализа трафика, а также распределенные системы обнаружения атак в промышленном интернете вещей, использующие параллельную обработку данных и обученные модели ML. Разрабатываются многоагентные подходы, где интеллектуальные агенты на базе методов глубокого обучения обменяются между собой знаниями об атаках без разглашения исходных данных (федеративный принцип). В работах Тушкановой и авторов исследованы подходы обнаружения кибератак и аномалий в киберфизических системах с различными источниками данных, приведена сравнительная оценка методов и показано, что сочетание символьных и нейросетевых методов улучшает выявление инцидентов на промышленных установках [15].

В особое направление следует выделить применение федеративного обучения для задач безопасности. Федеративное обучение позволяет нескольким организациям или узлам обучать общую модель обнаружения атак, не передавая друг другу сырые данные (важно для сохранения конфиденциальности. В работах авторов предложена архитектура системы обнаружения вторжений на основе федеративного ML и проведены эксперименты, подтвердившие эффективность такого подхода [16]. Архитектура включает специальные компоненты для выборки локальных данных, обучения локальных моделей, оценивания рисков утечки информации и выявления атак на сам процесс федеративного обучения. Результаты показывают, что модели IDS, обученные федеративно на распределенных данных разных организаций, обнаруживают атаки с точностью порядка 90–92%, что срав-

нимо с централизованным обучением, при существенном сокращении объема передаваемой информации. Однако в таком подходе присутствует ряд недостатков, таких как: отсутствуют единые методические рекомендации по построению и оценке таких систем; требуются механизмы противодействия возможным атакам на сами федеративные схемы (например, отравление данных участниками). Тем не менее, федеративный подход рассматривается как перспективный для сложных технических систем, где обмен сырыми данными между сегментами нежелателен, а совместная защита необходима.

Экспериментальное сравнение методов. Для сравнения рассмотренных методов проведен эксперимент с применением ML/DL к сетевому датасету UNSW-NB15. Выбранный набор данных содержит сбалансированную выборку нормального трафика и различных атак, что позволяет оценить эффективность алгоритмов на многоклассовой задаче. Обучение моделей Random Forest (RF) и глубокой нейронной сети проводилось на тренировочной выборке; затем производилась классификация тестовых данных. В качестве базовой линии для сравнения условно рассматривался сигнатурный метод, способный обнаруживать только атаки известного типа. На рис. 1 приведено сравнение интегральной точности обнаружения для традиционного сигнатурного подхода, алгоритма ML (Random Forest), модели DL (нейросеть) и их гибридного сочетания. Очевидно, что ML и DL существенно превосходят сигнатурную систему по точности (более 90% против ~70%). Модель глубокого обучения достигает около 96% точности, а гибридная многоуровневая система (комбинирующая сигнатурный и ML-анализ) – до ~98%, показывая, что сочетание разных методов позволяет повысить качество обнаружения.

Для глубокого нейросетевого классификатора рассчитаны показатели обнаружения по каждому классу атак. Рис. 2 иллюстрирует полноту обнаружения (recall) для 10 категорий трафика UNSW-NB15. Чаще встречающиеся типы атак, такие как Generic (массовые сетевые атаки), DoS и Reconnaissance (разведка) распознаются моделью практически полностью (recall ~95–99%). Для редких классов (Shellcode, Worms), на которые приходится мало обучающих примеров, полнота ниже (70–75%). Такие результаты отражают проблему несбалансированных данных: модель хуже выявляет малочисленные атаки [17]. В практических системах этот эффект смягчается методами балансировки, что частично учтено при тренировке нейросети. Тем не менее, даже редкие атаки («shellcode», «черви») обнаруживаются с приемлемым качеством (~70% обнаружения), тогда как без использования DL они зачастую остаются незамеченными.

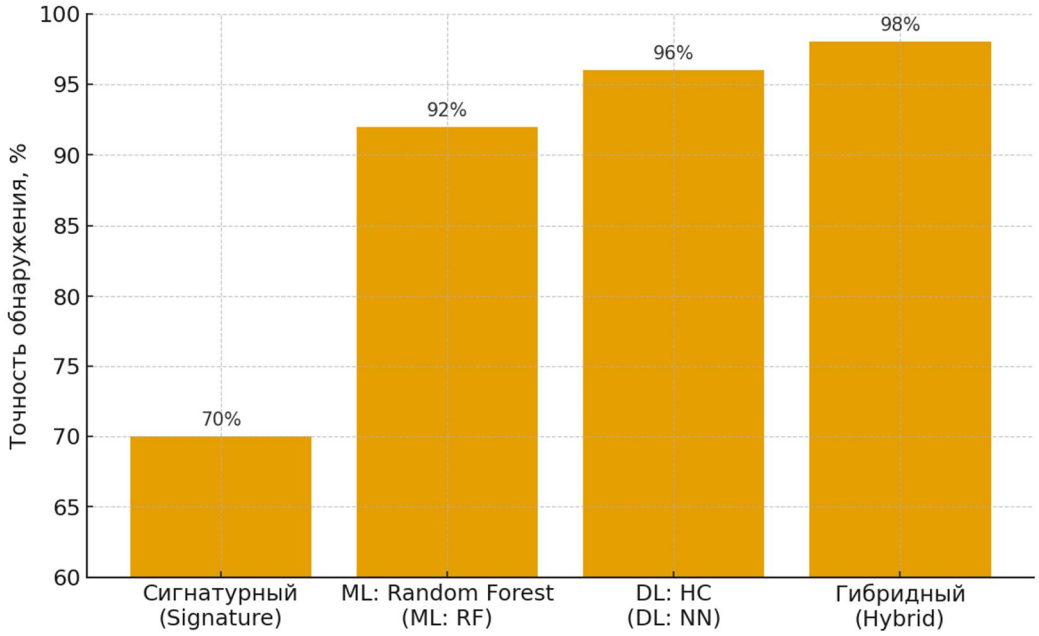

Рис. 1. Сравнение точности обнаружения атак различными подходами (сигнатурный метод, алгоритм ML, модель DL и гибридный ансамбль)

Fig. 1. Comparison of attack detection accuracy by different approaches (signature-based method, ML algorithm, DL model, and hybrid ensemble)

Кроме того, проанализирована производительность централизованного и федеративного подходов к обнаружению атак [18, 19]. В централизованной IDS все данные собираются и обрабатываются в одном узле (например, в дата-центре), тогда как при федеративном обучении модели обновляются на местах и обмениваются только обобщёнными параметрами.

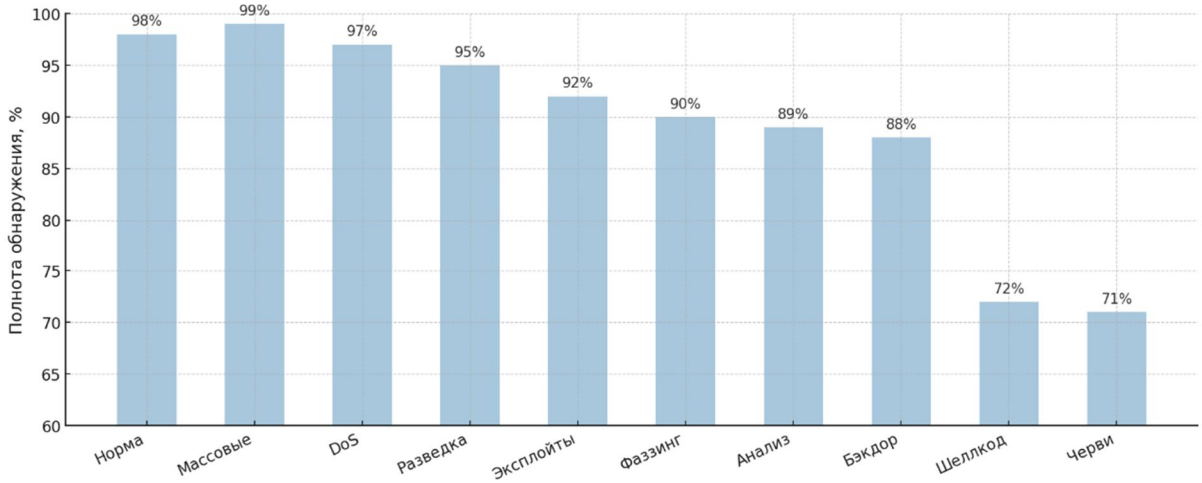

Рис. 2. Полнота обнаружения (Recall) по классам атак для модели глубокого обучения на датасете UNSW-NB15

Fig. 2. Per-class detection recall for the deep learning model on the UNSW-NB15 dataset

На рис. 3 приведено сравнительное диаграммное отображение точности обнаружения и средней задержки обработки событий для централизованной и федеративной IDS (по данным экспериментов из работы).

Из рис. 3 видно, что федеративная модель обеспечивает практически ту же точность обнаружения (~92%) что и централизованная (~95%), при этом средняя задержка обработки инцидентов сокращается примерно на 70% за счёт локальной обработки данных. Такой подход повышает оперативность реагирования и снимает нагрузку с центрального узла. Он особенно актуален для распределенных объектов (энергосети, сети связи, промышленные предприятия с филиалами), где объединение данных затруднено или нежелательно из-за требований безопасности.

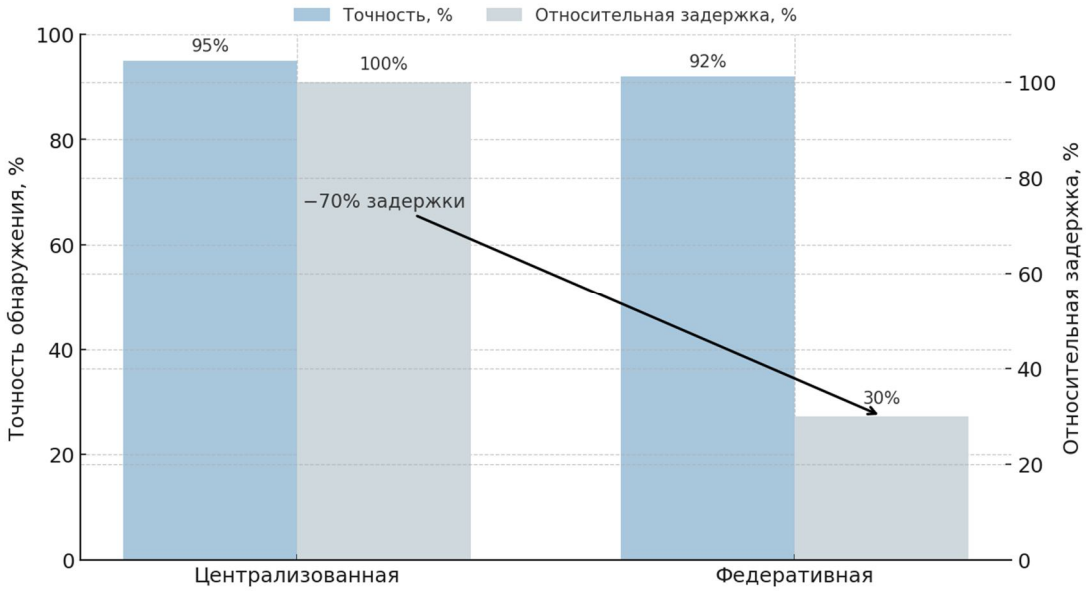

Рис. 3. Сравнение централизованной и федеративной модели IDS: точность обнаружения атак и относительная задержка обработки событий

Fig. 3. Comparison of centralized vs. federated IDS models: attack detection accuracy and relative event processing delay

Выводы

Искусственный интеллект является одним из наиболее перспективных направлений развития средств обнаружения и предотвращения деструктивных воздействий на информационные и технические системы. Проведенный анализ и эксперимент подтверждают, что интеграция алгоритмов ML/DL в системы кибербезопасности позволяет выявлять атаки с большей точностью и на более ранних этапах, чем традиционные сигнатурные методы. Особенно важно, что ИИ-средства способны распознавать новые и сложные угрозы, ранее ускользавшие от контроля, тем самым повышая ситуаци-

онную осведомленность операторов и сокращая «слепые зоны» в защите.

Полученные результаты подтверждают, что применение методов ИИ существенно повышает эффективность обнаружения атак по сравнению с традиционными средствами. Алгоритмы ML и DL способны обнаруживать известные и новые атаки с высокой точностью, сокращая долю пропущенных инцидентов. Глубокое обучение особенно эффективно для сложных сценариев атак за счет автоматического выявления скрытых аномалий. Однако без специальных мер DL-модели могут хуже детектировать редкие виды атак. Данная проблема требует внимания при развертывании систем (например, использование методов балансировки данных, генерации синтетических примеров атак и т.п.). Кроме того, следует учитывать вычислительные затраты: нейросетевые модели требуют более мощного аппаратного обеспечения и оптимизации, особенно для работы в реальном времени. В проведенном эксперименте подтвердилась перспектива применения ансамблей и гибридных систем: сочетание правил и обучаемых моделей дает более высокие показатели, объединяя достоинства детерминированных и вероятностных подходов.

Практическая ценность полученных результатов состоит в том, что разработанные подходы и рекомендации могут быть применены при создании новых и модернизации существующих систем кибербезопасности на сложных технических объектах различной инфраструктуры (энергетика, телекоммуникации, промышленный IoT и др.). Внедрение интеллектуальных агентов обнаружения на разных уровнях управления повысит способность инфраструктуры противостоять современным атакам. Важно отметить, что дальнейшие исследования должны быть направлены на повышение устойчивости ИИ-моделей. Требуется проведение более широких испытаний методов ML/DL в натурных условиях (на полигонах с имитацией работы реальной инфраструктуры) для оценки масштабируемости и надежности решений. Кроме того, необходимо развивать стандартизацию применения ИИ в кибербезопасности: отраслевые руководства и стандарты (в перспективе – ГОСТ) по использованию ML/DL для обнаружения атак существенно облегчат практическое внедрение. Перспективным направлением является внедрение концепции цифрового двойника безопасности, то есть создание модели защищаемой системы, на которой можно прорабатывать сценарии атак и упреждающе настраивать механизмы защиты. Решение обозначенных задач будет способствовать созданию проактивных, адаптивных и живучих киберсистем, способных противостоять самым современным и сложным атакам. Это является залогом безопасности как отдельных организаций, так и государства в целом в условиях цифровой эпохи.

Список литературы

1. Подсистема предупреждения компьютерных атак на объекты критической информационной инфраструктуры Российской Федерации / И.В. Котенко, А.И. Колесников, И.Б. Саенко, Р.И. Захарченко, Д.В. Величко // Вопросы кибербезопасности. 2023. № 1(53). С. 13–27. https://doi.org/10.21681/2311-3456-2023-1-13-27.

2. Advancing cybersecurity: a comprehensive review of AI-driven detection techniques / A.H. Salem, S.M. Azzam, O.E. Emam, A.A. Abohany // Journal of Big Data. 2024. Vol. 11, no. 1. P. 1–38. https://doi.org/10.1186/s40537-024-00957-y.

3. Jensen B., Atalan Y., Macias J.M. Algorithmic Stability: How AI Could Shape the Future of Deterrence // Center for Strategic and International Studies (CSIS). 2024. URL: https://www.csis.org/analysis/algorithmic-stability-how-ai-could-shape-future-deterrence.

4. Deep learning for intrusion detection in emerging technologies: a comprehensive survey and new perspectives / E.C. Pinto Neto, S. Iqbal, S. Buffett, M. Sultana, A. Taylor // Artificial Intelligence Review. 2025. Vol. 58. Art. 340. https://doi.org/10.1007/s10462-025-11346-z.

5. Pang G., Shen C., Cao L., van den Hengel A. Deep learning for anomaly detection: challenges, methods and opportunities. Preprint: arXiv:2007.02500, 2020. URL: https://arxiv.org/abs/2007.02500.

6. Sowmya T., Mary Anita E.A. A comprehensive review of AI-based intrusion detection system // Measurement: Sensors. 2023. Vol. 28. Article 100827. https://doi.org/10.1016/j.measen.2023.100827.

7. Issa M.M., Aljanabi M., Muhialdeen H.M. Systematic literature review on intrusion detection systems: research trends and future directions (2018–2023) // Journal of Intelligent Systems. 2024. (Early access). https://doi.org/10.1515/jisys-2023-0248.

8. Zhang Y., Muniyandi R.C., Qamar F. A Review of Deep Learning Applications in Intrusion Detection Systems: Overcoming Challenges in Spatiotemporal Feature Extraction and Data Imbalance // Applied Sciences. 2025. Vol. 15, no. 3. Art. 1552. https://doi.org/10.3390/app15031552.

9. APT Attack Detection Based on Graph Convolutional Neural Networks / W. Ren, X. Song, Y. Hong, Y. Lei, J. Yao, Y. Du, W. Li // International Journal of Computational Intelligence Systems. 2023. Vol. 16. Art. 184. https://doi.org/10.1007/s44196-023-00369-5.

10. Костогрызов А.И. Прогнозирование рисков по данным мониторинга для систем искусственного интеллекта // БИТ. Сборник трудов Десятой международной научно-технической конференции. М.: МГТУ им. Н. Э. Баумана, 2019. С. 220 – 229.

11. Kumar G., Thakur K., Ayyagari M.R. MLEsIDSs: machine learning-based ensembles for intrusion detection systems – a review // Journal of Supercomputing. 2020. Vol. 76, no. 12. P. 8938–8971. https://doi.org/10.1007/s11227-020-03196-z.

12. Bobyr M.V., Milostnaya N.A., Bulatnikov V.A. The fuzzy filter based on the method of areas' ratio // Applied Soft Computing. 2022. Vol. 117. Art. 108449. https://doi.org/10.1016/j.asoc.2022.108449.

13. The Current Research Status of AI-Based Network Security Situational Awareness / M. Wang, G. Song, Y. Yu, B. Zhang // Electronics. 2023. Vol. 12, no. 10. Art. 2309. https://doi.org/10.3390/electronics12102309.

14. Котенко И.В., Израилов К.Е., Буйневич М.В. Метод обнаружения атак различного генеза на сложные объекты на основе информации состояния. Ч. 1 // Вопросы кибербезопасности. 2023. № 3(55). С. 90–100. https://doi.org/10.21681/2311-3456-2023-3-90-100.

15. Detection of Cyberattacks and Anomalies in Cyber-Physical Systems: Approaches, Data Sources, Evaluation / O. Tushkanova, D. Levshun, A. Branitskiy, E. Fedorchenko, E. Novikova, I. Kotenko // Algorithms. 2023. 16(2). P. 85. https://doi.org/10.3390/a16020085.

16. Обнаружение вторжений на основе федеративного обучения: архитектура системы и эксперименты / Е.С. Новикова, И.В. Котенко, А.В. Мелешко, К.Е. Израилов // Вопросы кибербезопасности. 2023. № 6(58). С. 50–66. https://doi.org/10.21681/2311-3456-2023-6-50-66.

17. Shanmugam V., Razavi-Far R., Hallaji E. Addressing Class Imbalance in Intrusion Detection: A Comprehensive Evaluation of Machine Learning Approaches // Electronics. 2025. 14(1): 69. https://doi.org/10.3390/electronics14010069.

18. Survey of federated learning in intrusion detection / H. Zhang, J. Ye, W. Huang, X. Liu, J. Gu // Journal of Parallel and Distributed Computing. 2024. Vol. 195: 104976. https://doi.org/10.1016/j.jpdc.2024.104976.

19. Израилов К.Е., Буйневич М.В. Метод обнаружения атак различного генеза на сложные объекты на основе информации состояния. Ч. 2. Алгоритм, модель и эксперимент // Вопросы кибербезопасности. 2023. № 4(56). С. 80–93. https://doi.org/10.21681/2311-3456-2023-4-80-93.

References

1. Kotenko I.V., Kolesnikov A.I., Saenko I.B., Zakharchenko R.I., Velichko D.V. Subsystem of prevention of computer attacks on objects of critical information infrastructure of the Russian Federation. *Voprosy kiberbezopasnosti = Cybersecurity issues*. 2023;(1):13–27. (In Russ.). https://doi.org/10.21681/2311-3456-2023-1-13-27

2. Salem A.H., Azzam S.M., Emam O.E., Abohany A.A. Advancing cybersecurity: a comprehensive review of AI-driven detection techniques. *Journal of Big Data*. 2024;11(1):1–38. https://doi.org/10.1186/s40537-024-00957-y

3. Jensen B., Atalan Y., Macias J.M. Algorithmic Stability: How AI Could Shape the Future of Deterrence. *Center for Strategic and International Studies (CSIS)*. 2024 Jun 10.

Available from: https://www.csis.org/analysis/algorithmic-stability-how-ai-could-shape-future-deterrence

4. Pinto Neto E.C., Iqbal S., Buffett S., Sultana M., Taylor A. Deep learning for intrusion detection in emerging technologies: a comprehensive survey and new perspectives. *Artificial Intelligence Review*. 2025;58:340. https://doi.org/10.1007/s10462-025-11346-z

5. Pang G., Shen C., Cao L., van den Hengel A. Deep learning for anomaly detection: challenges, methods and opportunities. Preprint: arXiv:2007.02500; 2020. Available from: https://arxiv.org/abs/2007.02500

6. Sowmya T., Mary Anita E.A. A comprehensive review of AI-based intrusion detection system. *Measurement: Sensors*. 2023;28:100827. https://doi.org/10.1016/j.measen.2023.100827

7. Issa M.M., Aljanabi M., Muhialdeen H.M. Systematic literature review on intrusion detection systems: research trends and future directions (2018–2023). *Journal of Intelligent Systems*. 2024;(early access). https://doi.org/10.1515/jisys-2023-0248

8. Zhang Y., Muniyandi R.C., Qamar F. A review of deep learning applications in intrusion detection systems: overcoming challenges in spatiotemporal feature extraction and data imbalance. *Applied Sciences*. 2025;15(3):1552. https://doi.org/10.3390/app15031552

9. Ren W., Song X., Hong Y., Lei Y., Yao J., Du Y., Li W. APT attack detection based on graph convolutional neural networks. *International Journal of Computational Intelligence Systems.* 2023;16:184. https://doi.org/10.1007/s44196-023-00369-5

10. Kostogryzov A.I. Forecasting risks based on monitoring data for artificial intelligence systems. In: *BIT. Sbornik trudov Desyatoi mezhdunarodnoi nauchno-tekhnicheskoi konferentsii = BIT. Proceedings of the Tenth International Scientific and Technical Conference*. Moscow: Bauman Moscow State Technical University; 2019. P. 220–229. (In Russ.)

11. Kumar G., Thakur K., Ayyagari M.R. MLEsIDSs: machine learning-based ensembles for intrusion detection systems – a review. *The Journal of Supercomputing*. 2020;76(12):8938–8971. https://doi.org/10.1007/s11227-020-03196-z

12. Bobyr M.V., Milostnaya N.A., Bulatnikov V.A. The fuzzy filter based on the method of areas' ratio. *Applied Soft Computing*. 2022;117:108449. https://doi.org/10.1016/j.asoc.2022.108449

13. Wang M., Song G., Yu Y., Zhang B. The current research status of AI-based network security situational awareness. *Electronics*. 2023;12(10):2309. https://doi.org/10.3390/electronics12102309

14. Kotenko I.V., Izrailov K.E., Buinevich M.V. A method for detecting attacks of various origins on complex objects based on state information. Part 1. *Voprosy kiberbezopasnosti = Cybersecurity issues*. 2023;3(55):90–100. (In Russ.). https://doi.org/10.21681/2311-3456-2023-3-90-100

15. Tushkanova O., Levshun D., Branitskiy A., Fedorchenko E., Novikova E., Kotenko I. Detection of cyberattacks and anomalies in cyber-physical systems: approaches, data sources, evaluation. *Algorithms*. 2023;16(2):85. https://doi.org/10.3390/a16020085

16. Novikova E.S., Kotenko I.V., Meleshko A.V., Izrailov K.E. Intrusion detection based on federated learning: system architecture and experiments. *Voprosy kiberbezopasnosti = Cybersecurity issues*. 2023;(6):50–66. (In Russ.). https://doi.org/10.21681/2311-3456-2023-6-50-66

17. Shanmugam V., Razavi-Far R., Hallaji E. Addressing class imbalance in intrusion detection: a comprehensive evaluation of machine learning approaches. *Electronics*. 2025;14(1):69. https://doi.org/10.3390/electronics14010069

18. Zhang H., Ye J., Huang W., Liu X., Gu J. Survey of federated learning in intrusion detection. *Journal of Parallel and Distributed Computing*. 2024;195:104976. https://doi.org/10.1016/j.jpdc.2024.104976

19. Izrailov K.E., Buinevich M.V. A method for detecting attacks of various origins on complex objects based on state information. Part 2. Algorithm, model and experiment. *Voprosy kiberbezopasnosti = Cybersecurity issues*. 2023;(4):80–93. (In Russ.). https://doi.org/10.21681/2311-3456-2023-4-80-93

Информация об авторах / Information about the Authors

Селиверстов Дмитрий Евгеньевич, кандидат технических наук, младший научный сотрудник, Институт проблем управления им. В. А. Трапезникова Российской академии наук», г. Москва, Российская Федерация, e-mail: Seliverstov_dmitriyy@rambler.ru, ORCID: http://orcid.org/0009-0004-8412-7873

Русаков Константин Дмитриевич, научный сотрудник, Институт проблем управления им. В. А. Трапезникова Российской академии наук, г. Москва, Российская Федерация, e-mail: rusakov@ipu.ru, ORCID: http://orcid.org/0009-0004-8412-7873

Dmitry E. Seliverstov, Cand. of Sci. (Engineering), Junior Researcher, V. A. Trapeznikov Institute of Control Sciences of the Russian Academy of Sciences, Moscow, Russian Federation, e-mail: Seliverstov_dmitriyy@rambler.ru, ORCID: http://orcid.org/0009-0004-8412-7873

Konstantin D. Rusakov, Researcher, V. A. Trapeznikov Institute of Control Sciences of the Russian Academy of Sciences, Moscow, Russian Federation, e-mail: rusakov@ipu.ru, ORCID: http://orcid.org/0009-0004-8412-7873