DOI 10.36622/1682-7813.2025.28.2.010
УДК 621.372.004

АТАКУЕМЫЕ СРЕДСТВА МАШИННОГО ОБУЧЕНИЯ: МЕТОДИКА ФОРМИРОВАНИЯ СЦЕНАРИЕВ И РИСК-АНАЛИЗА ПРОЦЕССОВ ИХ РЕАЛИЗАЦИИ

Д.А. Нархов, К.В. Козина, Д.В. Кульшин, М.Д. Неменущий

В статье рассматривается процесс формирования множества сценариев атак на средства машинного обучения с использованием графовой модели, построенной на основе данных о реальных инцидентах информационной безопасности. Сформированный граф отражает взаимосвязи между техниками атак, уязвимостями программного обеспечения средств машинного обучения и условиями их эксплуатации. На основании структурного анализа графа выделены группы типовых сценариев атак. Для каждой группы проведён детальный риск-анализ. При оценке критичности уязвимостей применён комплексный подход с использованием специализированных калькуляторов: CVSS 3.1, CVSS 4.0, EPSS и AIVSS, что обеспечило многоаспектный анализ. Особое внимание уделено расчёту показателей риска: определены предварительные условия для перехода от одной техники к другой в рамках сценария атаки, вычислены произведения вероятностей таких переходов и суммарные величины потенциального ущерба. Результатом исследования стало выявление наиболее критичных сценариев с высокой мерой риска реализации.

Ключевые слова: средства машинного обучения, сценарии атак, уязвимости, риск-анализ, риск.

Введение

В первом квартале 2025 года российское киберпространство столкнулось с беспрецедентными вызовами информационной безопасности, обусловленными экспоненциальным ростом внедрения средств машинного обучения (СМО). Согласно данным Министерства цифрового развития [1], уровень проникновения решений с элементами СМО в коммерческом секторе достиг 54,3%, а в государственных учреждениях – 38,7%. При этом динамика роста продолжает ускоряться: если в 2021–2023 годах среднегодовой прирост составлял 7–9 процентных пунктов, то за первый квартал 2025 года показатель увеличился на 3,1 процентных пункта.

Государственный сектор демонстрирует особую активность в области внедрения технологий машинного обучения (МО). По данным Росстата [2], 72% федеральных органов исполнительной власти уже используют системы на основе МО для обработки документооборота, 58% – для аналитики больших данных, а 43% – для поддержки управленческих решений. В коммерческом секторе лидерами внедрения стали финансовые организации (89%), телекоммуникационные компании (76%) и ритейл (68%).

Однако бурное развитие технологий СМО сопровождается ростом связанных с ними киберугроз. Согласно исследованиям Positive Technologies [3-4] и Group-IB [5], в 2024 году количество атак на ИИ-системы увеличилось на 37% по сравнению с предыдущим годом, причем наиболее распространенными были атаки типа adversarial examples (27%), data poisoning (23%) и model inversion (18%). В первом же квартале 2025 года тенденция роста атак увеличилась на 12% по сравнению с предыдущим кварталом.

На фоне активного роста внедрения технологий СМО особенно остро встаёт проблема отсутствия нормативно-правовой базы и современных, адекватных методических подходов к идентификации атак на МО-системы и эффективному реагированию на них.

Анализ показывает следующую картину, что Доктрина информационной безопасности Российской Федерации, утвержденная Указом Президента № 646 от 5

декабря 2016 года [6], не содержит прямых положений, посвященных МО. Хотя в документе упоминаются перспективные информационные технологии, включая искусственный интеллект (ИИ), конкретные механизмы защиты систем МО не прописаны.

Частичное регулирование вопросов, связанных с машинным обучением, можно найти в следующих нормативных актах.

Приказ ФСТЭК России № 21 от 14.03.2013 [7] регламентирует требования по защите информации в государственных информационных системах, но не учитывает специфику МО в данном вопросе.

Указ Президента РФ № 124 от 15.02.2024 "О развитии искусственного интеллекта в Российской Федерации" [8] содержит общие принципы развития ИИ, однако не затрагивает вопросы обеспечения информационной безопасности МО. Стратегия развития ИИ до 2030 года [9] делает акцент на технологическом прогрессе и внедрении ИИ в различные сферы, при этом оставляя без внимания аспекты противодействия атакам на системы МО и минимизации последствий их реализации.

Также, современные методические подходы к идентификации атак на СМО и реагированию на них остаются далекими от совершенства. В частности, система MITRE ATLAS [10], предназначенная для систематизации знаний о тактиках и техниках атак на модели МО, обладает рядом существенных ограничений. Среди них – неполный охват актуальных векторов угроз, среднее отставание в обновлении базы данных новых атак на 4–6 месяцев, а также недостаточная проработка методических рекомендаций по обеспечению безопасности для конкретных платформ, архитектур и фреймворков. Кроме того, существенным недостатком данной платформы является отсутствие сформированных групп сценариев атак на МО.

Построение множества сценариев атак на средства МО, представляющего собой комплексную «карту» взаимосвязей между техниками, позволяет более точно реконструировать возможные траектории действий злоумышленника в информационных автоматизированных системах (ИАС), включающих компоненты МО. Такая систематизация обеспечивает специалистам по информационной безопасности возможность прогнозирования потенциальных векторов вторжения, а также своевременного выявления и нейтрализации угроз на ранних этапах их реализации – до достижения критически значимых целей и нанесения ущерба информационной инфраструктуре.

Таким образом, актуальность исследования обусловлена следующими факторами:

- стремительным ростом внедрения технологий МО в коммерческом и государственном секторах Российской Федерации при одновременном фрагментарном характере существующих подходов к идентификации атак на компоненты МО в составе ИАС, включая отсутствие систематизированного множества сценариев атак, необходимых для их эффективного предупреждения и предотвращения;
- необходимостью системного управления рисками в ИАС с МО и отсутствием специализированной методики их оценки.

Целью исследования является: повышение защищенности ИАС, включающих в себя СМО, за счет формирования множества сценариев атак на СМО и риск-анализа их реализации.

Для достижения поставленной цели необходимо выполнить следующие задачи:

- построить графовую модель множества сценариев атак на СМО;
- сформировать группы сценариев атак на МО и провести риск-анализ наиболее критичных сценариев.

Формирование множества сценариев атак на средства машинного обучения

В результате данного исследования были проанализированы атаки на средства машинного обучения, в том числе использовалось специально разработанное программное обеспечение для поиска уязвимостей, которые может эксплуатировать злоумышленник на каждом этапе реализации техники атаки. Удалось выделить двадцать шесть групп сценариев атак, а далее было определено несколько

тысяч конкретных сценариев атак для каждой группы. На основе данной работы построен невзвешенный граф множества сценариев атак на СМО, представленный на рис. 1.

Ниже приведены двадцать шесть групп сценариев атак на СМО, сформированных на основе реальных инцидентов информационной безопасности и включающих в себя техники, изображенные на графе (рис. 1).

Группа сценариев атаки №1 «Компрометация LLM-ассистентов для финансового обмана через инъекцию контента и манипуляцию результатами поиска»:

T0006 → T0049 → T0035→ T0055 → T0025→ T0010.003.

Группа сценариев атаки №2 «Обход обнаружения C&C-трафика через атаки на модель глубокого обучения средствами состязательных примеров»:

T0049→ T0042→ T0015.

Группа сценариев атаки №3 «Манипулирование DGA-доменами для обхода систем на основе МО»:

T0002 → T0042 → T0015.

Группа сценариев атаки №4 «Манипуляция обучающими данными в системах анализа вредоносного ПО»:

T0016.000 → T0043 → T0010.002 → T0020.

Группа сценариев атаки №5 «Зеркальное проникновение: подмена лиц в системах аутентификации»:

T0016.001 → T0016.000 → T0021→ T0047 → T0015.

Группа сценариев атаки №6 «Манипуляции и репликации моделей машинного перевода через API и публичные данные»:

T0002.000 → T0040 → T0005.001 → T0015 → T0031.

Группа сценариев атаки №7: «Нарушение безопасности ML-инфраструктуры из-за неправильной настройки и контроля доступа»:

T0021 → T0036 → T0002 → T0031.

Группа сценариев атаки №8 «Атаки на модели фильтрации спама с использованием прокси-моделей»:

T0047 → T0005.001→ T0015.

Группа сценариев атаки №9 «Атаки на целостность моделей разговорных агентов через обучение с подкреплением вредоносным контентом»:

T0047 → T0010.002 → T0020 → T0031.

Группа сценариев атаки №10 «Комбинированная атака на модель ИИ с доступом к API и прямой эксфильтрацией»:

T0012 → T0035 → T0025 → T0043.000 → T0040 → T0042 → T0015.

Группа сценариев атаки №11 «Модификация моделей ИИ в мобильных приложениях с внедрением управляющих бэкдоров»:

T0018.001 → T0042 → T0010.003 → T0043.004 → T0041 → T0015.

Группа сценариев атаки №12 «Оптические атаки на модели идентификации в физической среде»:

T0012 → T0040 → T0002.000 → T0043.000 → T0041 → T0015.

Группа сценариев атаки №13 «Обход облачных моделей ИИ в антивирусной защите»:

T0001 → T0047 → T0002.000 → T0042 → T0015.

Группа сценариев атаки №14 «Злоупотребление доверенными источниками в поставках ML-библиотек»:

T0010.001 → T0037 → T0025.

Группа сценариев атаки №15 «Компрометация LLM-приложения через инъекцию подсказок и выполнение сгенерированного кода»:

T0001 → T0047 → T0042 → T0049 → T0053 → T0055 → T0029.

Группа сценариев атаки №16 «Исполнение в облаке: атака через интерактивные окружения»:

T0010.001 → T0012 → T0011 → T0035 → T0025 → T0048.

Группа сценариев атаки №17 «Компрометация цепочки поставок моделей искусственного интеллекта через отравление весов LLM»:

T0018.000 → T0042 → T0058 → T0010.003 → T0031.

Группа сценариев атаки №18 «Атаки через поддельные зависимости и ложные пакеты, вызванные галлюцинациями больших языковых моделей»:

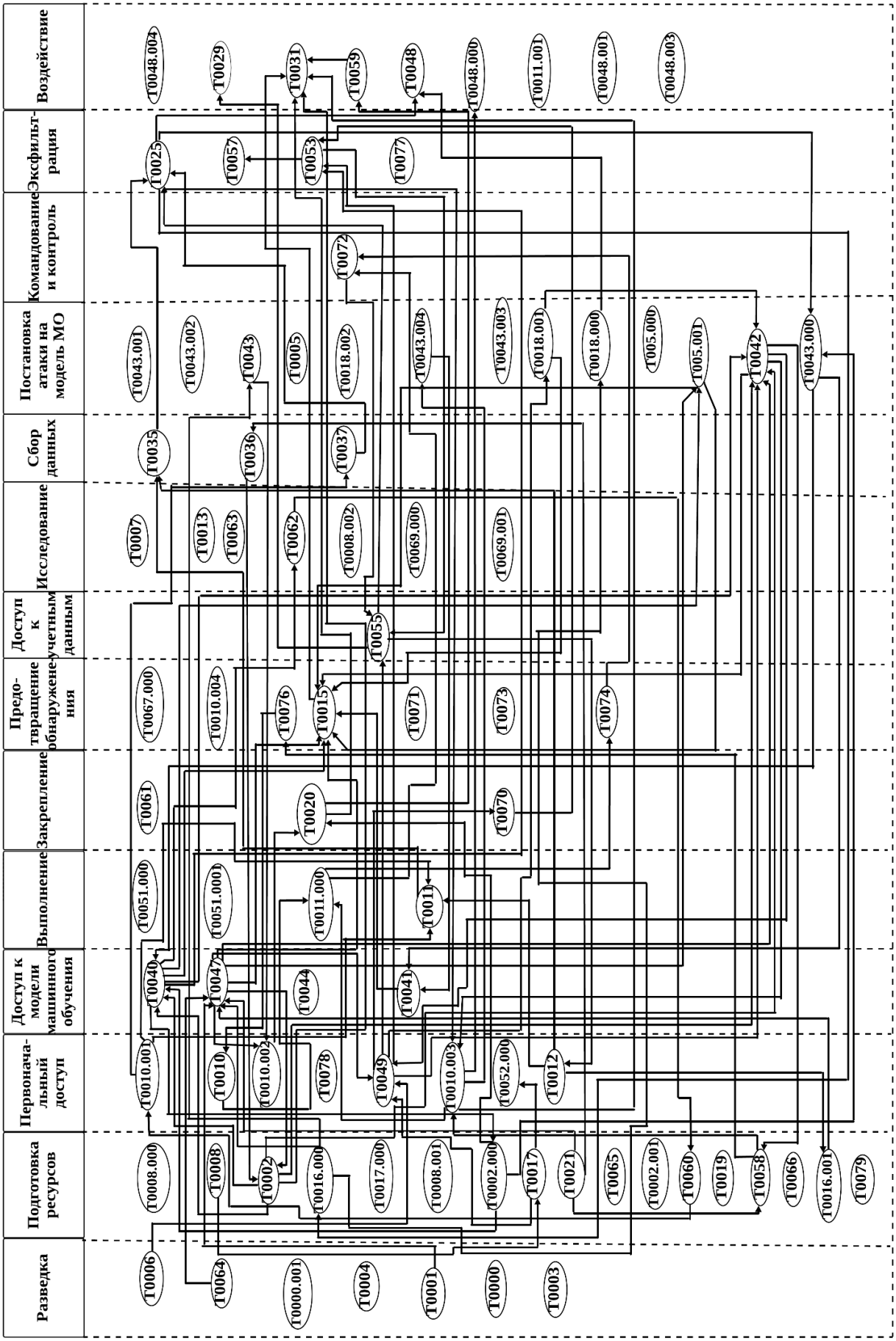

Рис. 1. Граф множества сценариев

T0040 → T0062 → T0060 → T0010.001 → T0011.001.

Группа сценариев атаки №19 «Самораспространяющиеся атаки на системы генерации с расширенным контекстом (RAG)»:

T0040 → T0053 → T0057.

Группа сценариев атаки №20 «Отравление данных в масштабе сети: атаки с разделением представления»:

T0002.000 → T0020 → T0059 → T0031.

Группа сценариев атаки № 21 «Неправомерный доступ к подключаемым модулям через инжектированный промпт»:

T0064 → T0047 → T0049 → T0070 → T0053.

Группа сценариев атаки № 22 «Атака через поддельные учетные записи в репозиториях моделей»:

T0021 → T0058 → T0010.003 → T0011.000 → T0074 → T0072 → T0055 → T0025 → T0016.000 → T0018.000 → T0048.

Группа сценариев атаки № 23 «Атака на жизненный цикл модели МО через уязвимые контейнерные реестры»:

T0049 → T0018.000 → T0018.001 → T0015.

Группа сценариев атаки № 24 «Непрямая инъекция подсказок и эксфильтрация данных через доверенные каналы визуализации:

T0008 → T0017 → T0049.

Группа сценариев атаки № 25 «Кража и монетизация LLM-ресурсов через облачные среды»:

T0049 → T0055 → T0012 → T0016.001.

Группа сценариев атаки № 26 «Эксплуатация десериализации моделей ИИ для удалённого выполнения кода»:

T0058 → T0076 → T0010 → T0011.000 → T0072.

Также определим предусловия злоумышленника, для наиболее характерных примеров выделенных сценариев, которые он должен получить в рамках действий одной техники, чтобы иметь возможность перейти в другую (табл. 1).

Оценка риска реализации сценария атак на средства машинного обучения

Так как в результате исследования в выделенных группах насчитываются тысячи сценариев реализации атак, рассмотрим наиболее показательные примеры, демонстрирующие эффективность разработанного методического обеспечения.

Риск-оценка соответствующих выделенных сценариев атак производится с использованием ниже представленного математического аппарата.

Оценка вероятности успешной реализации уязвимости по формуле (1).

$$p_{\text{ij}} = \frac{\overline{\text{k}_\text{i}}}{\sum_{j=1}^{m} \overline{k_{ij}}} \quad (1)$$

где p_{ij} – вероятность реализации атаки i-м вектором на j-ую уязвимость,

k_i – оценка критичности i-й уязвимости,

n – общее количество уязвимостей.

Оценка ущерба атаки вычисляется по формуле (2).

$$u_{ij} = 1 - (1 - u_{\text{к}})(1 - u_{\text{д}})(1 - u_{\text{ц}}), \quad (2)$$

где u_{ij} – ущерб реализации атаки i-м вектором на j-ую уязвимость,

$u_{\text{к}}$, $u_{\text{ц}}$, $u_{\text{д}}$ – воздействие на конфиденциальность, целостность, доступность.

Из оценок (1) и (2) риск рассчитывается по формуле (3).

$$Risk_{ij} = \text{p}_{\text{ij}} u_{ij} \quad (3)$$

Оценка критичности уязвимости выстраивается с помощью использования четырех калькуляторов: CVSS 3.1 [11], CVSS 4.0 [12], EPSS [13], AIVSS [14]. Если первые три калькулятора являются достаточно распространенными и универсальными для оценки любых уязвимостей, то калькулятор AIVSS используется конкретно для оценки уязвимостей средств машинного обучения и учитывает специфику данной направленности, вводя специализированные метрики, например: MR (надежность модели), DS (чувствительность данных), DC (критичность решения), LL (уязвимости жизненного цикла), AA (подверженность методам состязательных атак).

Таблица 1

Предварительные условия для перехода в технику

№ группы сценария	Предварительные условия	Техника	Последствия
5	Наличие среды разработки, виртуальной машины, SDK/IDE и offensive ML-фреймворков.	T0016.001: Подготовка программных средств: программные средства	Настроенная тестовая среда с возможностью эмуляции атак на модели.
	Настроенная тестовая среда с возможностью эмуляции атак на модели.	T0016.000: Подготовка необходимых программных средств: готовые реализации атак на модели	Наличие инструментов для генерации adversarial-контента (deepfake, morphing).
	Наличие инструментов для генерации adversarial-контента.	T0021: Создание учётных записей	Доступ к активным ML-учётным записям, привязанным к украденным персональным данным.
	Доступ к активным ML-учётным записям.	T0047: Доступ к продукту или сервису, использующему МО	Возможность отправки запросов в ML-сервис (через API или пользовательский интерфейс) от имени жертвы.
	Возможность отправки запросов в ML-сервис (через API или пользовательский интерфейс) от имени жертвы.	T0015: Обход модели МО	Успешное внедрение поддельной личности и доступ к защищённой инфраструктуре через обман ML-модели.
11	Злоумышленник обладает ресурсами и возможностью разместить (использовать) автономное оборудование для обучения моделей.	T0018.001: Получение технической инфраструктуры: собственное оборудование	Получена автономная инфраструктура для подготовки и модификации модели без контроля со стороны третьих лиц.
	Получен полный контроль над инфраструктурой, позволяющей обучать модели и внедрять модифицированные данные в процессе обучения.	T0043.004: Создание состязательных данных: вставка бэкдор триггера	Сформированы обучающие данные с встроенными триггерами, предназначенными для активации непредсказуемого поведения модели.
	Имеется возможность обучить или дообучить модель с использованием бэкдорных данных и подготовить её к публикации или распространению.	T0010.003: Компрометация цепочки поставок МО: модели	Модифицированная модель загружена в цепочку поставок (например, в репозиторий), доступную конечным пользователям.
	Модифицированная модель попадает в окружение, где обрабатываются реальные пользовательские данные, включая те, что поступают из физической среды (аудио, видео, сенсоры).	T0041: Доступ к данным, поступающим из физической среды	Модель получает и обрабатывает реальные данные, содержащие потенциальные триггеры, активирующие вредоносное поведение.
	Обнаружено, что данные из физической среды вызывают активацию триггера и обход стандартной логики работы модели.	T0015: Обход модели МО	Активация скрытого поведения модели, проявляющегося при наличии бэкдор-триггера, нарушать ожидаемую функциональность модели.
	Подготовлена среда для проведения испытаний, обеспечена возможность наблюдать поведение модели при заданных условиях.	T0042: Подтверждение эффективности атаки	Подтверждено, что внедренный бэкдор работает корректно, активация происходит при заданных входных условиях, готовность к эксплуатации.

Продолжение табл.1

№ группы сценария	Предварительные условия	Техника	Последствия
17	Доступ к публичному репозиторию моделей ИИ и возможность загрузки исходной модели.	T0018.000: Создание бэкдора в модели машинного обучения: обучение модели на отравленных данных	Отравленная модель содержит ложную информацию и демонстрирует деструктивное поведение.
	Отравленная модель с ложной (вредоносной) информацией доступна для тестирования.	T0042: Подтверждение эффективности атаки	Успешная атака подтверждена, модель эффективно подменена и незаметна.
	Подтверждённая работоспособность отравленной модели.	T0058: Публикация отравленных моделей	Отравленная модель опубликована в общедоступном репозитории под именем, похожим на оригинал.
	Доступ к опубликованной отравленной модели для загрузки.	T0010.003: Компрометация цепочки поставок МО: модели	Пользователи загружают и интегрируют отравленную модель в свои приложения и системы.
	Интеграция отравленной модели в пользовательские системы и приложения.	T0031: Нарушение целостности модели МО	Генерируются вредоносные данные, что приводит к компрометации выходов модели, деградации её работоспособности и снижению доверия к модели со стороны пользователей.

Для группы сценариев №5 определен наиболее критичный сценарий атаки: T0016.001(CVE-2023-23770) → T0016.000(CVE-2024-48700) → T0021(CVE-2019-8950) → T0047(CVE-2018-17190) → T0015(CVE-2016-1516). Модель сценария продемонстрирована на рис. 2.

Проведем оценку критичности выявленных уязвимостей (табл. 2), а следующим этапом на основании данных из табл. 2 рассчитаем вероятность эксплуатации данного сценария, сумму ущербов и риск (табл. 3).

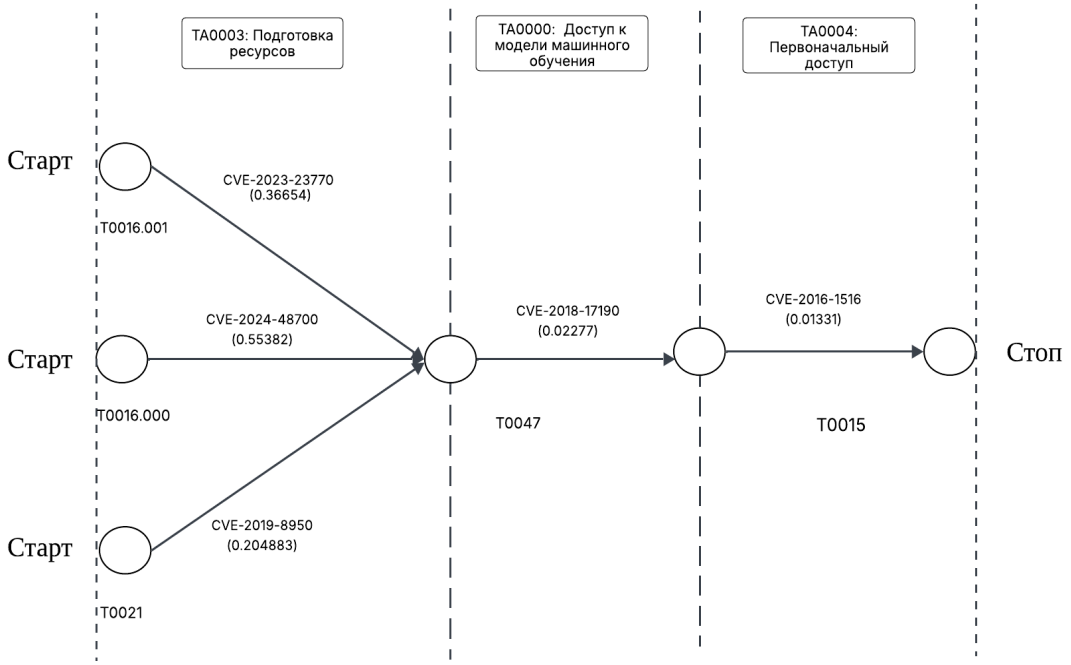

Рис. 2. Модель сценария №5

Таблица 2

Критичность уязвимостей для сценария атаки №5

Уязвимость	CVSS 3.1	CVSS 4.0	EPSS	AIVSS	Среднеквадратичная
CVE-2023-23770	9.4	8.8	0.85	0.5122090625	0.875755281669195
CVE-2024-48700	7.2	8.6	0.131	0.1664919375	0.738250682063576
CVE-2019-8950	9.8	9.3	0.892	0.2763370625	0.926525442473465
CVE-2018-17190	9.8	9.3	0.855	0.2472190625	0.91655593588743
CVE-2016-1516	8.8	8.6	0.999	0.223186375	0.909891052578627

Таблица 3

Оценка риска наиболее критичного сценария №5

Произведение вероятностей	Сумма ущербов	Риск
0,0000136498252155686	4.3000	0,0000586942484269451

Для пятой группы было сформировано 2001 сценарий, из этих сценариев приведен пример реализации с наибольшим риском, $Risk$ = 0,000058694248426945.

Рассмотрим наиболее критичный сценарий для группы сценариев № 11: T0018.001(CVE-2022-32985) → T0042(CVE-2025-3121) → T0010.003(CVE-2020-28593) → T0043.004(CVE-2025-25362) → T0041(CVE-2023-42143) → T0015(CVE-2016-1516). Модель сценария изображена на рисунке 3, расчеты представлены в табл. 4 и 5. Для одиннадцатой группы наиболее критичный сценарий реализации атаки имеет риск $Risk$ = 0,007466.

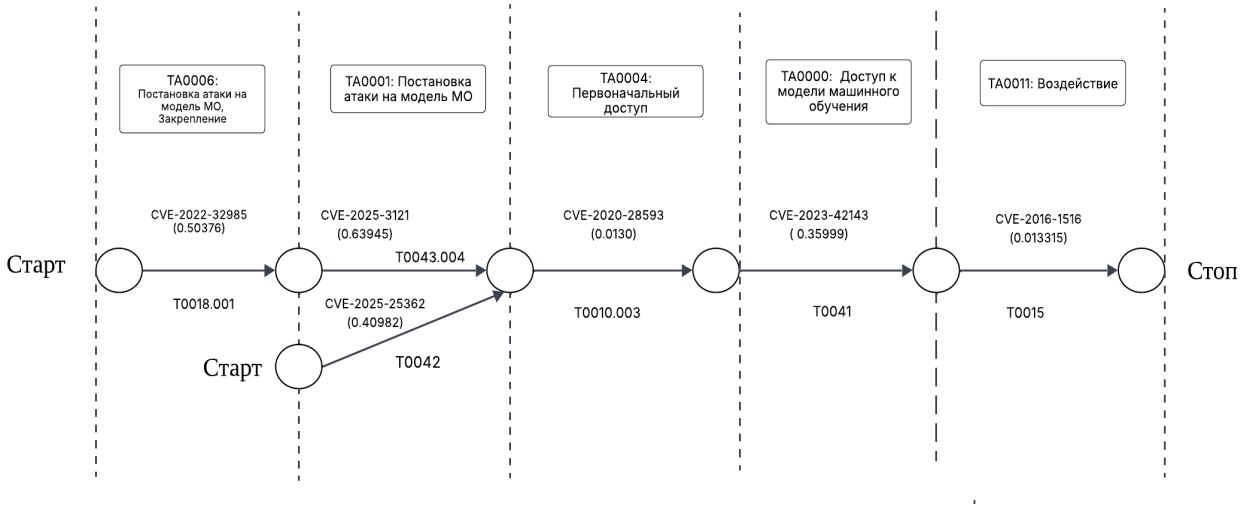

Рис. 3. Модель сценария № 11

Таблица 4

Критичность уязвимостей для сценария атаки №11

Уязвимость	CVSS 3.1	CVSS 4.0	EPSS	AIVSS	Среднеквадратичная
CVE-2022-32985	9.8	9.3	0.473	0.2330290625	0.851751464920002
CVE-2025-3121	3.3	4.8	0.19	0.262886125	0.366457433675535
CVE-2020-28593	8.1	9.2	0.1387	0.30932375	0.801573252731672
CVE-2025-25362	9.8	9.3	0.114	0.3252490625	0.678113904430253

Продолжение табл.4

Уязвимость	CVSS 3.1	CVSS 4.0	EPSS	AIVSS	Среднеквадратичная
CVE-2023-42143	5.4	4.8	0.1	0.189302175	0.467419640777594
CVE-2016-1516	8.8	8.6	0.999	0.223186375	0.909891052578627

Таблица 5

Оценка риска наиболее критичного сценария №11

Произведение вероятностей	Сумма ущербов	Риск
0,001839	4.0600	0,007466

Для группы сценариев №17 наиболее критичный сценарий с высоким риском: T0018.000(CVE-2024-1880) → T0042(CVE-2025-3121) → T0058(CVE-2019-14281) → T0010.003(CVE-2020-28593) → T0031(CVE-2023-34362) (рис. 4). Расчеты представлены в табл. 6 и 7. Для семнадцатой группы, имеющей 2100 сценариев, наиболее критичный сценарий реализации атаки имеет риск $Risk$ = 0,014234.

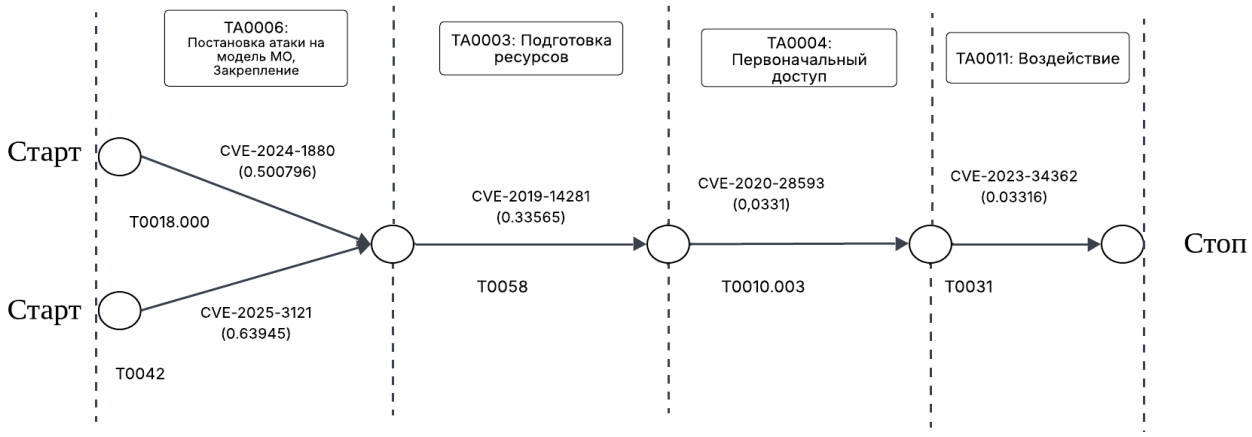

Рис. 4. Модель критичного сценария для группы №17

Таблица 6

Критичность уязвимостей для сценария атаки №17

Уязвимость	CVSS 3.1	CVSS 4.0	EPSS	AIVSS	Среднеквадратичная
CVE-2024-1880	7.8	8.5	0.304	0.223208125	0.727802009554432
CVE-2025-3121	3.3	4.8	0.19	0.262886125	0.366457433675535
CVE-2019-14281	9.8	9.3	0.1216	0.2516170625	0.894960104453477
CVE-2020-28593	8.1	9.2	0.1387	0.30932375	0.801573252731672
CVE-2023-34362	9.8	9.3	0.94485	0.2478190625	0.944098327107978

Таблица 7

Оценка риска наиболее критичного сценария №17

Произведение вероятностей	Сумма ущербов	Риск
0,003847	3.7000	0,014234

Для группы сценариев №22 наиболее критичный сценарий с высоким риском: T0021(CVE-2019-8950) → T0058(CVE-2019-14281) → T0010.003(CVE-2020-28593) → T0011.000(CVE-2024-6960) → T0074(CVE-2022-40425) → T0072(CVE-2022-40809) →

T0055(CVE-2023-38896) → T0025(CVE-2023-6014) → T0016.000(CVE-2024-48700) → T0018.000(CVE-2024-1880) → T0048(CVE-2024-37061). Модель сценария представлена на рис. 5. Расчеты представлены в табл. 8 и 9.

Для двадцать второй группы, насчитывающей 2254 сценария, наиболее критичный сценарий реализации атаки имеет риск $Risk$ = 0.0000028208054971886466.

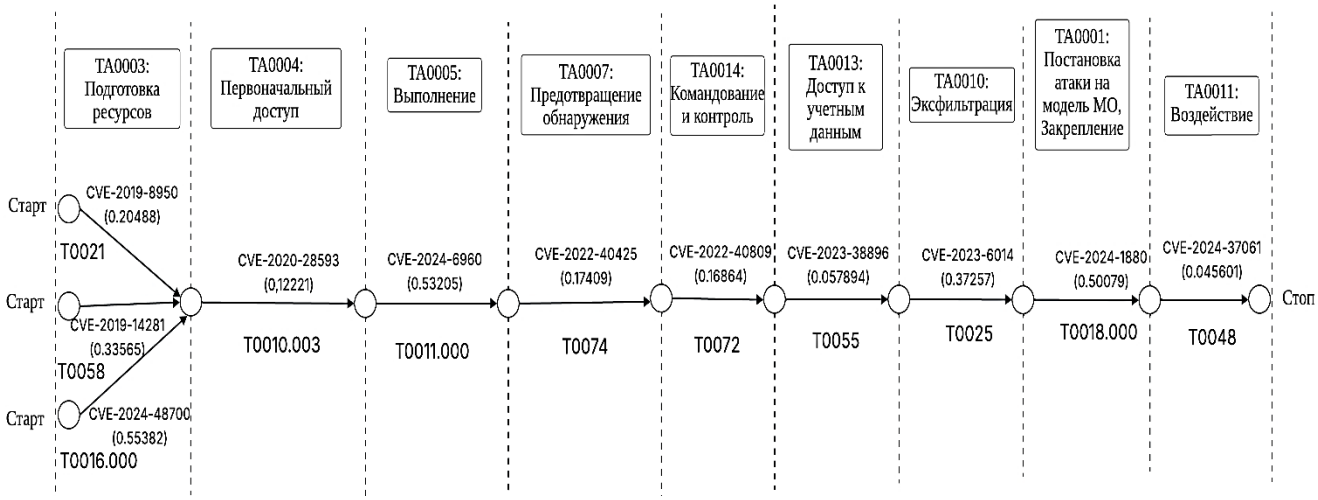

Рис. 5. Модель критичного сценария для группы №22

Таблица 8

Критичность уязвимостей для сценария атаки №22

Уязвимость	CVSS 3.1	CVSS 4.0	EPSS	AIVSS	Среднеквадратичная
CVE-2019-8950	9.8	9.3	0.892	0.2763370625	0.926525442473465
CVE-2019-14281	9.8	9.3	0.1216	0.2516170625	0.894960104453477
CVE-2020-28593	8.1	9.2	0.1387	0.30932375	0.801573252731672
CVE-2024-6960	7.5	7.5	0.99	0.2088445	0.838563542918832
CVE-2022-40425	9.8	9.3	0.97	0.2369290625	0.952842274694858
CVE-2022-40809	9.8	9.3	0.142	0.3421690625	0.885159539806755
CVE-2023-38896	9.8	9.3	0.788	0.2330290625	0.899123364701583
CVE-2023-6014	9.8	9.3	0.842	0.2130730625	0.913969395164871
CVE-2024-48700	7.2	8.6	0.131	0.1664919375	0.738250682063576
CVE-2024-1880	7.8	8.5	0.304	0.223208125	0.727802009554432
CVE-2024-37061	8.8	8.6	0.958	0.245800375	0.893405572927441

Таблица 9

Оценка риска наиболее критичного сценария №22

Произведение вероятностей	Сумма ущербов	Риск
0,0000003234	9.6250	0.0000028208054971886466

Таким образом, из сформированных групп сценариев были выделены наиболее критичные сценарии, имеющие высокий риск реализации.

Заключение

В настоящем исследовании были достигнуты поставленные задачи: построен невзвешенный граф множества сценариев атак на средства машинного обучения, представляющий собой эффективный инструмент для анализа возможных векторов атак. Данный граф позволяет наглядно представить, каким образом злоумышленник может выстроить сценарий атаки, какие техники и уязвимости могут быть использованы, а также определить необходимые предусловия для перехода от одной техники к другой. Это, в свою очередь, способствует более глубокому пониманию механизмов компрометации систем машинного обучения. На основании полученного графа выполнена кластеризация сценариев по их признаковому сходству, что позволило выделить типовые группы атак.

Дополнительно проведена количественная оценка рисков реализации нескольких сценариев атак. Наиболее критичные сценарии, обладающие высоким уровнем риска, были выделены и проанализированы. Расчет критичности уязвимостей производился с применением специализированного калькулятора оценки уязвимостей, предназначенного для анализа уязвимостей, специфичных для средств машинного обучения.

Новизна результатов, полученных в работе, состоит в том, что:

- впервые построено множество сценариев атак на СМО;
- впервые выделены группы сценариев атак на СМО и определены наиболее критичные сценарии реализации атаки по каждой группе.

Полученные результаты представляют собой ценную методическую основу для специалистов в области информационной безопасности. Используя предложенную методику, можно воспроизводимо и обоснованно формировать комплекс мер по обеспечению устойчивости систем машинного обучения к современным угрозам, а также организовать эффективное противодействие возможным атакам.

Список литературы

1. 2024 Информационно-аналитическая справка по результатам мониторинга внедрения решений в сфере искусственного интеллекта в приоритетных отраслях экономики Российской Федерации по итогам 1-го полугодия 2024 года, НЦРИИ URL: https://ai.gov.ru/knowledgebase/vnedrenie-ii/2023_informacionno analiticheskaya_spravka_po_rezulytatam_monitoringa_vnedreniya_resheniy_v_sfere_iskusstvennogo_intellekta_v_prioritetnyh_otraslyah_ekonomiki_rossiyskoy_federacii_po_itogam_2023_goda_ncrii/ (дата обращения 23.05.25).

2. Федеральная служба государственной статистики URL: https://rosstat.gov.ru/ (дата обращения 23.05.25).

3. Отчет об Индексе ИИ URL: https://ai.gov.ru/knowledgebase/infrastruktura-ii/2024_otchet_ob_indekse_ii_artificial_intelligence_index_report_2024_stanford/ (дата обращения 23.05.25).

4. Актуальные киберугрозы: III квартал 2024 года URL: https://www.ptsecurity.com/ru-ru/research/analytics/aktualnye-kiberugrozy-iii-kvartal-2024-goda/ (дата обращения 23.05.25).

5. Group-IB URL: https://www.group ib.com/products/business-email protection (дата обращения 23.05.25).

6. Указ Президента РФ от 5 декабря 2016 г. N 646 «Об утверждении Доктрины информационной безопасности Российской Федерации» URL: https://base.garant.ru/71556224/ (дата обращения 23.05.25).

7. Приказ ФСТЭК России № 21 от 14.03.2013 (редакция от 14.05.2020) «Об утверждении состава и содержания организационных и технических мер по обеспечению безопасности персональных данных при их обработке в информационных системах персональных данных» URL: https://fstec.ru/dokumenty/vse-dokumenty/prikazy/prikaz-fstek-rossii-ot-18-fevralya-2013-g-n-21 (дата обращения 23.05.25).

8. Указ Президента РФ № 124 от 15.02.2024 «О развитии искусственного интеллекта в Российской Федерации» URL: http://www.kremlin.ru/acts/bank/50326 (дата обращения 23.05.25).

9. Национальная стратегия развития ИИ на период до 2030 года утверждена Указом Президента РФ от 10.10.2019 №490. URL:

http://www.kremlin.ru/acts/bank/44731 (дата обращения 23.05.25).

10. METRE ATLAS URL: https://atlas.mitre.org/ (дата обращения 23.05.2025).

11. Common Vulnerability Scoring System v3.1 URL: https://www.first.org/cvss/v3.1/ (дата обращения 23.05.25).

12. Common Vulnerability Scoring System v4.0 URL: https://www.first.org/cvss/v4.0/ (дата обращения 23.05.25).

13. EPS-Calculator URL: https://theowni.github.io/EPSS-Calculator/ (дата обращения 23.05.25).

14. OWASP Artificial Intelligence Vulnerability Scoring System URL: https://owasp.org/www-project-artificial-intelligence-vulnerability-scoring-syste (дата обращения 23.05.2025).

Воронежский государственный технический университет
Voronezh State Technical University

Поступила в редакцию 26.05.25

Информация об авторах

Нархов Дмитрий Андреевич – аспирант, кафедра систем информационной безопасности, Воронежский государственный технический университет, e-mail: alexanderostapenkoias@gmail.com

Козина Ксения Владимировна – студентка, кафедра систем информационной безопасности, Воронежский государственный технический университет, e-mail: alexanderostapenkoias@gmail.com

Кульшин Дмитрий Вячеславович – студент, кафедра систем информационной безопасности, Воронежский государственный технический университет, e-mail: alexanderostapenkoias@gmail.com

Неменущий Максим Дмитриевич – студент, кафедра систем информационной безопасности, Воронежский государственный технический университет, e-mail: alexanderostapenkoias@gmail.com

ATTACKED MACHINE LEARNING TOOLS: THE FORMATION OF MULTIPLE SCENARIOS AND RISK ANALYSIS OF THEIR IMPLEMENTATION PROCESSES

D.A. Narhov, K.V. Kozina, D.M. Kulshin, M.D. Nemenushchiy

The article examines the process of forming multiple attack scenarios for machine learning tools using a graph model based on data on real information security incidents. The generated graph reflects the interrelationships between attack techniques, machine learning software vulnerabilities, and their operating conditions. Based on the structural analysis of the graph, groups of typical attack scenarios are identified. A detailed risk analysis was carried out for each group. When assessing the criticality of vulnerabilities, an integrated approach was applied using specialized calculators: CVSS 3.1, CVSS 4.0, EPSS and AIVSS, which provided a multidimensional analysis. Special attention is paid to the calculation of risk indicators: the prerequisites for the transition from one technique to another within the framework of the attack scenario are determined, the products of the probabilities of such transitions and the total values of potential damage are calculated. The result of the study was the identification of the most critical scenarios with a high risk of implementation.

Keywords: machine learning tools, attack scenarios, vulnerabilities, risk analysis, risk.

Submitted 26.05.25

Information about the authors

Dmitry A Narhov – postgraduate student, department of information security systems, Voronezh state technical University, e-mail: alexanderostapenkoias@gmail.com

Ksenia V. Kozina – student, department of information security systems, Voronezh state technical University, e-mail: alexanderostapenkoias@gmail.com

Dmitry V. Kulshin – student, department of information security systems, Voronezh state technical University, e-mail: alexanderostapenkoias@gmail.com

Maksim D. Nemenushchiy – student, department of information security systems, Voronezh state technical University, e-mail: alexanderostapenkoias@gmail.com