Научная статья
УДК 004.89
DOI: 10.26583/bit.2025.4.01

ФОРМИРОВАНИЕ РАЗМЕЧЕННОГО НАБОРА ДАННЫХ НА ОСНОВЕ СМОДЕЛИРОВАННЫХ КОМПЬЮТЕРНЫХ АТАК

Алексей В. Павлычев[1], Кирилл В. Кузьминец[2], Данил Е. Бреус[3], Александр А. Шелупанов[4]
[1-3]Дальневосточный федеральный университет (ДВФУ), п. Аякс, 10, Владивосток, 690922, Россия
[4]Томский государственный университет систем управления и радиоэлектроники (ТУСУР), Ленина пр-т, 40, Томск, 634050, Россия
[1]pavlychev.av@dvfu.ru, https://orcid.org/0009-0006-1779-7935
[2]kuzminetc.kv@dvfu.ru, https://orcid.org/0009-0006-2258-538X
[3]breus.de@dvfu.ru, https://orcid.org/0009-0007-5910-3425
[4]saa@tusur.ru, https://orcid.org/0000-0003-2393-6701

Аннотация. В рамках работы предложен подход к созданию массива данных, необходимого для будущих систем машинного обучения искусственного интеллекта, предназначенных для противодействия компьютерным атакам. Авторами проведен краткий обзор современных способов выявления компьютерных атак, в том числе с помощью анализа лог-файлов операционной системы с применением методов машинного обучения. Рассмотрены различные подходы к моделированию атак и формированию наборов данных. Предложен алгоритм для формирования базы данных из записей системных журналов и соответствующих им компьютерных атаках и сценариях пользовательской работы, учитывающий различные подходы и лучшие практики. Для реализации алгоритма спроектирована и развернута виртуальная сетевая инфраструктура, предназначенная для автоматизированного извлечения файлов системных журналов операционной системы после запуска заранее подготовленных скриптов. В результате проведенного исследования сформулирован алгоритм моделирования компьютерных атак и формирования размеченного набора данных из системных журналов операционной системы. Согласно разработанному алгоритму в результате проведенной серии экспериментов получена база данных, которую можно использовать для обучения моделей машинного обучения и применения в современных интеллектуальных средствах защиты информации. Датасет, содержащий 1 473 559 записей, размещен на платформе Kaggle. Научная новизна заключается в доработке алгоритма моделирования компьютерных атак и формирования набора данных из системных журналов операционной системы, который в отличие от существующих подходов автоматизирует процесс сбора данных, что обеспечивает более полное покрытие тактик MITRE ATT&CK. Усовершенствованный алгоритм включает механизмы валидации выполнения скриптов и передачи метаданных, что повышает достоверность данных, получаемых в ходе выполнения скриптов и формирования датасета, в отличие от алгоритмов, не учитывающих данные аспекты.

Ключевые слова: *алгоритм, компьютерные атаки, тактики и техники, системные журналы, машинное обучение, набор данных.*

Для цитирования: *Павлычев, Алексей В. и др. Формирование размеченного набора данных на основе смоделированных компьютерных атак. Безопасность информационных технологий, [S.l.], т. 32, № 4, с. 1–17, 2025. ISSN 2074-7136. URL: https://bit.spels.ru/index.php/bit/article/view/1858. DOI: http://dx.doi.org/10.26583/bit.2025.4.01.*

Scientific article

FORMATION OF A LABELED DATASET BASED ON SIMULATED CYBER ATTACKS

Aleksey V. Pavlychev[1], Kirill V. Kuzminetc, Danil E. Breus, Aleksander A. Shelupanov
[1-3]Far Eastern Federal University (FEFU), Ajax, 10, Vladivostok, 690922, Russia
[4]Tomsk State University of Control Systems and Radioelectronics (TUSUR), 40, Lenina Pr., Tomsk, 634050, Russia
[1]pavlychev.av@dvfu.ru, https://orcid.org/0009-0006-1779-7935
[2]kuzminetc.kv@dvfu.ru, https://orcid.org/0009-0006-2258-538X
[3]breus.de@dvfu.ru, https://orcid.org/0009-0007-5910-3425
[4]saa@tusur.ru, https://orcid.org/0000-0003-2393-6701

Abstract. This paper proposes an approach is proposed to create a data corpus necessary for future artificial intelligence and machine learning systems designed to counter computer attacks. The authors have provided a brief overview of modern methods for detecting computer attacks, including the analysis of the log files of the operating system using machine learning methods. Various approaches to attack modeling and dataset formation are considered. An algorithm is proposed for generating a database from system log entries and related computer attacks and user scenarios, taking into account various approaches and best practices. To implement the algorithm, a virtual network infrastructure has been designed and deployed, designed for automated extraction of operating system log files after running pre-prepared scripts. As a result of the study, an algorithm for modeling computer attacks and the formation of a labeled dataset from the system logs of the operating system is formulated. According to the developed algorithm, as a result of a series of experiments, a database has been obtained that can be used to train machine learning models and application in modern intelligent information security tools. The dataset containing 1,473,559 records is hosted on the Kaggle platform. The scientific novelty of the work lies in the refinement of the algorithm for modeling computer attacks and generating a dataset from the operating system logs. Unlike existing approaches, automates the data collection process, ensuring more comprehensive coverage of MITRE ATT&CK tactics. The improved algorithm includes mechanisms for validating script execution and metadata transmission, which increases the reliability of data obtained during script execution and dataset generation, compared to algorithms that do not account for these aspects.

Keywords: *algorithm, computer attacks, tactics and techniques, system logs, machine learning, dataset.*

For citation: *Pavlychev, Aleksey V. et al. Formation of a labeled dataset based on simulated cyber attacks. IT Security (Russia), [S.l.], v. 32, no. 4, p. 1–17, 2025. ISSN 2074-7136. URL: https://bit.spels.ru/index.php/bit/article/view/1858. DOI: http://dx.doi.org/10.26583/bit.2025.4.01.*

Введение

В условиях стремительной цифровизации всех сфер жизни общества проблема кибербезопасности становится одной из наиболее значимых. Компьютерные атаки представляют собой серьезную угрозу для государственных, корпоративных и частных информационных систем, что делает исследования в области их обнаружения и ликвидации крайне актуальными.

Традиционные системы защиты, такие как антивирусы и межсетевые экраны, зачастую не справляются с современными угрозами. Необходим поиск новых подходов к выявлению и ликвидации последствий современных кибератак.

Современные компьютерные атаки представляют собой сложные и многоаспектные угрозы, которые сочетают в себе различные объекты воздействия и инструменты для достижения своих целей. Они могут быть направлены на нарушение конфиденциальности, целостности или доступности данных, а также использовать как технические уязвимости, так и методы социальной инженерии. Атаки часто включают в себя несколько этапов, таких как разведка, внедрение, эксплуатация и сокрытие следов, что делает их сложными для обнаружения и предотвращения.

Кроме того, злоумышленники активно применяют современные технологии, включая искусственный интеллект и машинное обучение, для автоматизации атак и

повышения их эффективности. От специалистов по кибербезопасности требуется использование комплексных подходов, включающих не только традиционные методы защиты, но и передовые технологии анализа данных, мониторинга и прогнозирования угроз.

Киберугрозы эволюционируют с беспрецедентной скоростью, поэтому моделирование компьютерных атак играет ключевую роль в противодействии им. Путем симуляции реалистичных сценариев атак на основе фреймворков, таких как MITRE ATT&CK, специалисты могут заранее тестировать системы обнаружения, разрабатывать стратегии защиты и обучать модели машинного обучения на разнообразных данных. Это позволяет не только предсказывать и предотвращать атаки, но и минимизировать ущерб, адаптируя меры безопасности к новым тактикам угроз, без риска для реальных систем.

В данной работе совершенствуется алгоритм формирования базы данных из записей системных журналов, который включает в себя симуляции атак с учётом тактик MITRE ATT&CK и пользовательской активности. Соблюдение баланса классов критично для обучения точных моделей машинного обучения, способных выявлять угрозы в реальном времени и снижать ложные срабатывания. В отличие от существующих подходов, предложенный алгоритм фокусируется на автоматизированном сборе и валидации данных в виртуальной инфраструктуре, что делает его доступным для практического применения в информационной безопасности.

Объектом настоящего исследования выступают подсистемы машинного обучения искусственного интеллекта для систем противодействия компьютерным атакам. Предмет исследования – методы формирования массивов исходных данных (датасетов) из системных журналов операционных систем, необходимых для обучения и тестирования таких подсистем.

Целью работы является создание массива исходных данных, необходимого для будущих систем машинного обучения искусственного интеллекта.

Для достижения цели поставлены следующие задачи:

1. Провести обзор современных методов выявления компьютерных атак с использованием анализа лог-файлов и машинного обучения.
2. Рассмотреть существующие подходы к моделированию атак и формированию наборов данных.
3. Разработать алгоритм формирования базы данных из записей системных журналов с учётом различных сценариев атак и пользовательской активности.
4. Спроектировать и реализовать виртуальную сетевую инфраструктуру для автоматизированного сбора и обработки системных журналов.
5. Провести эксперименты по формированию датасета и оценить возможность его применения для обучения моделей машинного обучения.

Ожидаемые результаты включают разработанный алгоритм и сформированный размеченный датасет, который может быть использован специалистами в области информационной безопасности и искусственного интеллекта.

Научная и практическая значимость заключается в автоматизации процесса сбора данных из системных журналов, обеспечении более полного покрытия тактик MITRE ATT&CK и повышении достоверности данных за счёт валидации выполнения скриптов и передачи метаданных.

1. Методы выявления компьютерных атак

Для противодействия различным компьютерным атакам специалисты по обеспечению информационной безопасности используют следующие основные методы и инструменты:

1. Анализ сетевого трафика [1].
2. Анализ системных журналов (лог-файлов) [2]. При работе с системными журналами используются следующие методы анализа [3]:
   2.1. Ключевые слова и фильтры.
   2.2. Корреляция событий.
   2.3. Машинное обучение для анализа лог-файлов.
3. Поведенческий анализ.
4. Анализ уязвимостей.

Машинное обучение представляет собой метод анализа данных, который автоматизирует построение аналитических моделей. Это ветвь искусственного интеллекта, основанная на идее, что системы могут учиться на данных, выявлять закономерности и принимать решения с минимальным вмешательством человека.

Машинное обучение доказало свою эффективность в решении различных аналитических задач и все чаще используется для обнаружения угроз и автоматического устранения их, прежде чем они смогут нанести ущерб. Разработанные алгоритмы быстро сканирует большие объемы данных и анализирует их с помощью статистики.

В машинном обучении первостепенное значение имеет набор данных. Прежде чем проводить какой-либо анализ, исследователь должен сам тщательно изучить и обработать набор имеющихся данных. В методах машинного обучения нельзя напрямую использовать необработанные данные, такие как захват пакетов (pcap), NetFlow, записи лог-файлов.

Данные должны быть предварительно обработаны, чтобы их можно было использовать в популярных инструментах машинного обучения, таких как R, RapidMiner, специальные библиотеки в Python, например, Scikit-Learn. Исследователи, использующие анализ машинного обучения, должны понимать методологию сбора данных и методы, используемые при предварительной обработке данных [4].

Рассмотрим примеры наиболее распространенных наборов данных, используемых в машинном обучении при решении задач в области кибербезопасности [5].

The DARPA (Defense Advanced Research Project Agency – Агентство перспективных оборонных исследований) располагает двумя наборами данных, имеющих большую ценность для исследователей кибербезопасности. Набор данных DARPA 1998 г. и 1999 г. был разработан группой Cyber Systems and Technology лаборатории Линкольна Массачусетского технологического института (MIT/LL) [6,7].

KDD 1999 – еще один известный набор данных, который в основном используется исследователями кибербезопасности (набор был специально подготовлен в рамках соревнований по анализу данных – The Third International Knowledge Discovery and Data Mining Tools Competition [8]). Набор KDD 1999 имеет 41 подробный атрибут и очень похож на данные NetFlow. Авторы статьи [9] всесторонне изучили набор данных KDD 1999. Статистический анализ показал, что в наборе данных было огромное количество избыточных записей. Обнаружено, что 78% обучающей выборки и 75% тестовой выборки одинаковы. Эти врожденные проблемы делают KDD 1999 непригодным для использования, и поэтому предложен новый набор данных, названный NSL-KDD [10].

Исследователи методов обнаружения компьютерных атак часто используют одни и те же наборы данных, для оценки своих алгоритмов. Но эти наборы данных устарели и могут не содержать актуальных признаков современных видов атак. Поэтому в настоящее время остро стоит проблема формирования наборов данных для использования в машинном обучении для решения актуальных задач кибербезопасности [11].

2. Алгоритм моделирования компьютерных атак и формирования набора данных

Обнаружение компьютерных атак остается актуальной темой ввиду постоянного появления новых видов киберугроз. Одним из перспективных направлений в данной

области является применение искусственного интеллекта и методов машинного обучения (Machine Learning, ML). В [12–13] отмечено, что алгоритмы машинного обучения могут значительно повысить точность обнаружения атак, позволяя системам учиться на больших объемах данных.

Проблема сбора и разметки данных для систем кибербезопасности является активно развивающейся областью научных знаний [14]. Использование машинного обучения, автоматизация разметки, анализ аномалий и стандартизация данных становятся ключевыми аспектами для повышения эффективности и точности систем защиты.

Для решения задачи автоматизированного сбора данных и их разметки в соответствии с проводимыми компьютерными атаками проектируются и создаются элементы информационной инфраструктуры, которые могут быть как физическими, так и виртуальными [15].

Для моделирования компьютерных атак исследователи часто используют матрицу MITRE ATT&CK (Adversarial Tactics, Techniques, and Common Knowledge), которая представляет собой структурированную базу знаний о поведении злоумышленников, которая описывает тактики и техники, используемые в кибератаках [16]. Матрица MITRE ATT&CK для корпоративных сетей включает множество тактик и техник.

С помощью фильтров можно выделить техники, которые можно выявить путем анализа системных журналов операционной системы Windows:

1. Первоначальный доступ (Initial Access).
2. Выполнение (Execution).
3. Закрепление (Persistence)
4. Повышение привилегий (Privilege Escalation).
5. Предотвращение обнаружения (Defense Evasion).
6. Получение учётных данных (Credential Access).
7. Обнаружение (Discovery).
8. Боковое перемещение (Lateral Movement).
9. Сбор данных (Collection).
10. Эксфильтрация данных (Exfiltration).
11. Деструктивное воздействие (Impact).

В открытом доступе небольшое количество научных работ, дающих описание процесса моделирования компьютерных атак и формирования наборов данных для их обнаружения.

В ранних исследованиях, таких как [17] описываются аналитические методы работы с файлами журналов. С помощью запросов SQL авторы собирают данные и вычисляют среднее количество событий определённого типа в любое время дня для любого сервера или пользователя в наборе данных. Затем определяется среднее количество событий определённого типа и стандартное отклонение этих событий, превышение которого является признаком компьютерной атаки.

Задачей выявления компьютерных атак с помощью анализа журналов операционной системы занимается группа исследователей Эгейского университета Греции. В [18–20] подробно рассматриваются вопросы обнаружения признаков бокового перемеения в сети с помощью инструмента Sysmon. Автором разработан и опубликован собственный инструмент для анализа журналов Python_Evtx_Analyzer .

Исследователи из Бельгийской Королевской военной академии в [21] делают упор на анализе тактик различных APT-группировок и выявлении техник закрепления в инфраструктуре. Для эмуляции атак и создания реалистичных сценариев использовались инструменты Caldera, Atomic Red Team. Для создания пользовательского трафика использовался инструмент GHOSTS Framework. В статье описана настройка виртуальных сред с различными пользовательскими профилями для сбора данных, а также инструменты автоматизации. Авторами собран и опубликован собственный датасет.

В работе Цюрихского университета [22] проведен подробный анализ выявления некоторых техник проведения компьютерных атак как традиционным сигнатурным способом, так и с применением алгоритмов машинного обучения. Основным

инструментом в ходе исследования являлся Splunk. Автором использовались журналы, опубликованные в открытом доступе. В работе приведены результаты сигнатурного выявления атак с использованием системных журналов операционной системы, которые составили 91,65%.

Анализ подходов к моделированию атак и формированию датасета приведен в табл. 1.

Таблица 1. Подходы к моделированию атак и формированию датасетов

Автор, год публикации	Описание атак по MITRE ATT&CK, покрытие тактик	Источник данных	Описание моделирования атак, инструменты	Описание формирования датасета, инструменты	Дополнительная информация
John Dwyer, Traian Marius Truta, 2013	нет	Win Event Logs	частично, ручной метод	частично, SQL-запросы	
Christos Smiliotopoulos at al., 2022, 2023	да, боковое перемещение	Sysmon	частично, ручной метод	да, ETCExptool[1]	
Khaled Rahal, Arbia Riah, Thibault Debatty, 2024	да, закрепление	Win Event Logs, Sysmon	да, для атак Caldera, Atomic Red Team, для пользователей GHOSTS Framework[2]	нет	Автоматизация атак с помощью YAML-правил, Chocolatey для развертывания инфраструктуры, валидация выполнения скриптов
Layla Husselman, 2025	да, повышение привилегий, получение учетных данных, предотвращение обнаружения, боковое перемещение	Win Event Logs	нет, использовались публичные наборы данных	да, Splunk[3]	В работе приведены результаты сигнатурного выявления данных атак, которые составили 91,65%
Настоящая работа	да, повышение привилегий, закрепление, получение учетных данных, предотвращение обнаружения	Win Event Logs	да, для атак Caldera, Atomic Red Team, для пользователей скрипты Python	да, Elastic[4]	Автоматизация атак с помощью YAML-правил, валидация выполнения скриптов, передача метаданных об атаках и пользовательской активности

Как видно из табл. 1, большинство исследователей, сосредотачиваются на изучении конкретной одной тактики, ввиду трудоемкости проведения экспериментов и формирования наборов данных.

[1]Описание инструмента по ссылке URL: https:// github.com/ChristosSmiliotopoulos/evtx_To_CSV_ExportTool (дата обращения: 18.09.2025).
[2]Описание инструмента по ссылке URL: https://cmu-sei.github.io/GHOSTS/ (дата обращения: 18.09.2025).
[3]Описание инструмента по ссылке URL: https://www.splunk.com/ (дата обращения: 18.09.2025).
[4]Описание инструмента по ссылке URL: https://www.elastic.co/ (дата обращения: 18.09.2025).

Можно выделить следующие лучшие практики, применяемые при моделировании:

1. Структурирование по MITRE ATT&CK. Использование общепринятой таксономии для классификации атак.

2. Виртуализация. Применение виртуальных сред для безопасного моделирования атак.

3. Автоматизация. Разработка скриптов и конфигурационных файлов для автоматизации процесса сбора и разметки данных.

4. Метаданные. Включение дополнительной информации для точной разметки событий.

5. Реалистичность. Стремление к созданию датасетов, отражающих реальные условия.

6. Баланс классов. Учет несбалансированности данных в реальных условиях, а также при обучении моделей.

7. Документация. Детальное описание процесса создания датасета для воспроизводимости.

Проведенный анализ существующих подходов и лучших практик дает возможность сформулировать алгоритм моделирования компьютерных атак и формирования набора данных из записей журналов операционной системы. Визуальное представление алгоритма в виде блок-схемы представлено на рис. 1.

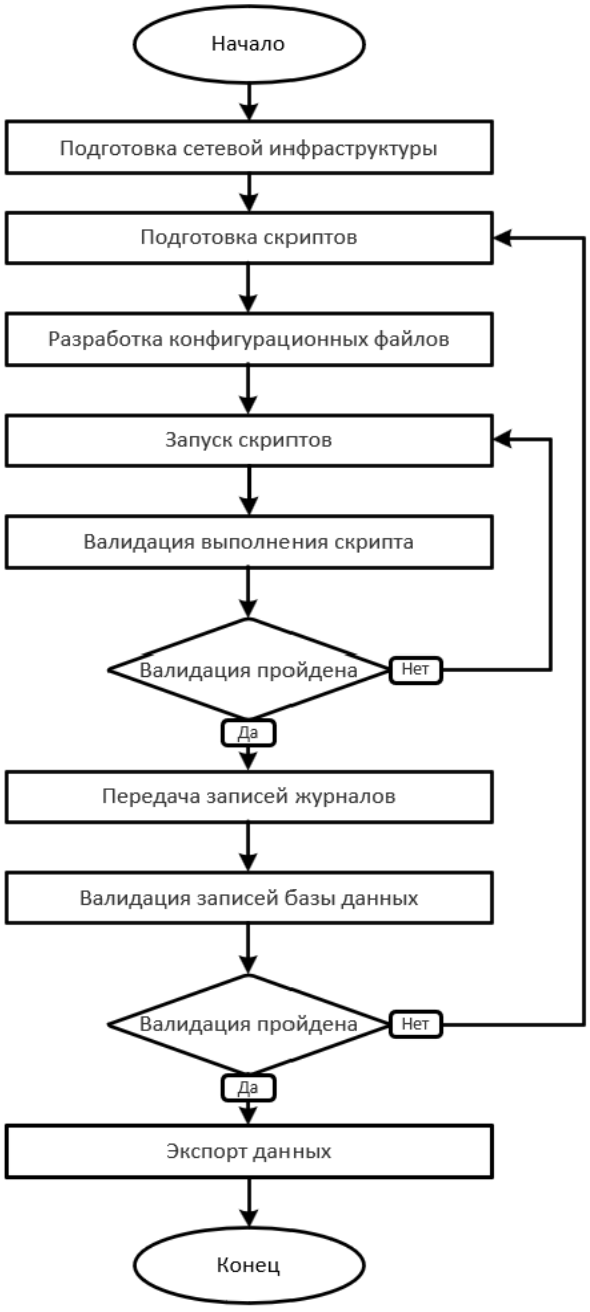

Рис. 1. Алгоритм формирования набора данных из записей системных журналов

Данный алгоритм включает в себя следующие шаги:

1. Подготовка набора скриптов атак и пользовательских действий. Атаки должны привязываться к матрице MITRE ATT&CK. Эмуляция пользовательских действий необходима для создания реалистичных данных для датасета.

2. Разработка конфигурационных файлов. Для автоматизации процесса формирования датасета и его воспроизводимости создаются конфигурационные YAML-правила – файлы, в которых описываются условия атаки (например, уязвимость, версия ПО), действия (эксплуатация уязвимости, выполнение команд, выполнение кода параметры (IP-адреса, порты), зависимости (например, необходимость предварительной аутентификации), а также количество итераций запуска скриптов. Инструменты

автоматизации (такие как Metasploit Framework) загружают эти правила и выполняют атаки в автоматическом или полуавтоматическом режиме.

3. Подготовка виртуальной инфраструктуры. На данном этапе готовится сетевая инфраструктура, на которой будут эмулироваться компьютерные атаки и легитимные действия пользователей. Также в инфраструктуре разворачиваются средства автоматизации и база данных для хранения полученных записей журналов операционной системы.

4. Запуск скриптов. Согласно загруженному конфигурационному файлу, осуществляется запуск скриптов, эмулирующих компьютерные атаки и легитимные действия пользователей.

5. Валидация выполнения скриптов. С помощью специальных скриптов осуществляется проверка выполнения скрипта (например, создание файла или запуск командной строки). В случае, если валидация не пройдена, проводится корректировка конфигурационного файла, осуществляется повторный запуск скриптов.

6. Передача записей системных журналов в базу данных. На данном этапе осуществляется передача и сохранение информации из системных журналов операционной системы. Также передаются метаданные (например, вид атаки или действия пользователя) и уникальный идентификатор сессии.

7. Валидация записей базы данных. На данном этапе проводится проверка полученных записей на предмет сбалансированности данных. Если данных недостаточно, либо фиксируется дисбаланс классов, с помощью конфигурационного файла проводятся дополнительные итерации запуска скриптов. Если в процессе валидации базы данных выявляются новые факторы или обнаруживаются ошибки или пробелы в данных, необходимо возвращаться к этапам подготовки скриптов для их корректировки, что позволит уточнять и дополнять датасет, учитывая новые условия и факторы атак.

8. Экспорт данных. На финальном этапе алгоритма осуществляется сохранение полученных данных в удобном для дальнейшего использования формате.

3. Реализация разработанного алгоритма

В рамках исследования спроектирован и развернут исследовательский стенд, содержащий следующие основные элементы:

Атакующий – виртуальная машина, с которой осуществляются компьютерные атаки;

Пользователь – виртуальная машина, на которую осуществляются компьютерные атаки. На ней также расположены разработанные скрипты, имитирующие легитимные действия пользователей. С данной машины пересылаются данные о записях системных журналов после каждой реализованной компьютерной атаки или пользовательского сценария;

Хранилище – база данных, в которую передаются записи из системных журналов пользовательской машины и метаданные, необходимые для разметки;

Управляющий скрипт – специально разработанный программный продукт, позволяющий в автоматическом режиме запускать компьютерные атаки и пользовательские сценарии, а также передавать метаданные о видах атак и действиях пользователей в хранилище для дальнейшей обработки.

В качестве программной среды для исследования выбрана следующая конфигурация: гипервизор Oracle VM VirtualBox[5] для эмуляции элементов компьютерной

[5]Описание инструмента по ссылке URL: https://www.virtualbox.org/ (дата обращения: 18.09.2025).

сети; виртуальная машина, на которой установлена ОС Kali Linux и специализированное ПО Caldera для эмуляции атак; виртуальная машина, на которой установлена ОС Windows 10, пользовательское программное обеспечение (офисный пакет, браузер, мессенджеры) и специализированное ПО Winlogbeat[6] для передачи системных журналов; виртуальная машина, на которой установлена ОС Debian и база данных Elastic для хранения информации из системных журналов и метаданных.

Действия, воспроизводимые в информационной системе в рамках исследования, включают в себя две группы:

1. Компьютерные атаки.
2. Легитимные действия пользователей.

В качестве программного средства для моделирования компьютерных атак использовался фреймворк Caldera с дополнительными плагинами. Также был разработан ряд скриптов, интегрированных в данный инструмент.

Для моделирования легитимных действий пользователей разработан скрипт на языке программирования Python, эмулирующий распространенные сценарии офисной работы:

1. Просмотр веб-страниц и видеоконтента при помощи браузера.
2. Работа с пакетом офисных приложений.
3. Получение и отправка электронной почты.
4. Использование файлового хранилища.

С целью автоматизации запуска компьютерных атак и различных пользовательских сценариев написан конфигурационный файл и разработан управляющий скрипт. Данный программный продукт также предназначен для передачи метаданных, для дальнейшей передачи в базу данных и разметки полученных результатов.

Передаваемые метаданные содержат информацию о виде действия: компьютерная атака (ATTACK), либо легитимные действия (USER_ACTION), а также номер техники из матрицы MITRE ATT&CK и конкретный тип пользовательских действий (WEB, OFFICE, FTP, MAIL).

Непосредственно процесс моделирования компьютерных атак и действий пользователей включает в себя следующие этапы:

1. Исследователь с использованием управляющего скрипта инициирует запуск различных компьютерных атак или пользовательских сценариев.

2. После завершения компьютерной атаки или пользовательского сценария осуществляется проверка успешности выполнения скрипта, журналы операционной системы передаются с использованием инструмента Winlogbeat в хранилище, где обрабатываются и сохраняются в базе данных.

3. После реализации каждой компьютерной атаки или пользовательского сценария пользовательская виртуальная машина сбрасывается в исходное состояние.

4. Из базы данных информация из системных журналов, а также метаданные экспортируются в удобном формате для дальнейшей обработки.

4. Описание полученного набора данных

В результате эксперимента сформирована база данных, содержащая размеченную информацию о записях системных журналов, полученных с виртуальной машины, на которой были реализованы различные компьютерные атаки и сценарии пользовательской

[6]Описание инструмента по ссылке URL: https://www.elastic.co/beats/winlogbeat (дата обращения: 18.09.2025).

работы. Общий вид полученной базы данных представлен на рис. 2. Итоговое количество собранных событий составило 1 473 559.

	@timestamp	winlog.channel	event.code	fields.category	fields.session	fields.sub_category
0	Apr 28, 2025 @ 15:11:40.080	Security	4624	USER_ACTION	20250427221113	OFFICE
1	Apr 28, 2025 @ 15:11:40.080	Security	4672	USER_ACTION	20250427221113	OFFICE
2	Apr 28, 2025 @ 15:11:40.048	System	6	USER_ACTION	20250427221113	OFFICE
3	Apr 28, 2025 @ 15:11:40.036	Security	4624	USER_ACTION	20250427221113	OFFICE
4	Apr 28, 2025 @ 15:11:40.036	Security	4672	USER_ACTION	20250427221113	OFFICE
...	...	...	...	...	...	...
1642041	May 1, 2025 @ 21:41:15.803	Security	4672	ATTACK	20250501214051	SYSTEM_BINARY_PROXY_EXECUTION_(T1218)
1642042	May 1, 2025 @ 21:41:15.799	Application	4625	ATTACK	20250501214051	SYSTEM_BINARY_PROXY_EXECUTION_(T1218)
1642043	May 1, 2025 @ 21:41:15.748	System	24	ATTACK	20250501214051	SYSTEM_BINARY_PROXY_EXECUTION_(T1218)
1642044	May 1, 2025 @ 21:41:15.748	System	1	ATTACK	20250501214051	SYSTEM_BINARY_PROXY_EXECUTION_(T1218)
1642045	May 1, 2025 @ 21:41:15.748	Security	4616	ATTACK	20250501214051	SYSTEM_BINARY_PROXY_EXECUTION_(T1218)

1473559 rows × 6 columns

Рис. 2. Визуальное представление сформированного набора данных

Описание полей в наборе данных:

1. @timestamp – дата и время регистрации события;
2. winlog.channel – системный журнал Windows, в котором зафиксировано событие. Принимает одно из следующих значений: Security, System, Application;
3. event.code – код события Event ID из набора Windows Event Logs;
4. fields.category – указывает на вид события и принимает два значения: ATTACK – компьютерная атака, USER_ACTION – пользовательские действия;
5. fields.session – уникальный идентификатор сессии, в рамках которого зафиксировано событие;
6. fields.sub_category – конкретный вид компьютерной атаки или пользовательского действия.

Компьютерные атаки маркируются согласно матрицы MITRE ATT&CK:

ACCOUNT_MANIPULATION_(T1098): тактика TA0004 Повышение привилегий (Privilege Escalation);

CREDENTIAL_BRUTEFORCE_(T1110): тактика TA0006 Получение учетных данных (Credential Access);

BITS_JOBS_(T1197): тактика TA0003 Закрепление (Persistence);

SYSTEM_BINARY_PROXY_EXECUTION_(T1218): тактика TA0005 Предотвращение обнаружения (Defense Evasion);

Виды пользовательских действий:

OFFICE – работа с офисными приложениями;

WEB – работа в Интернет-браузере;

MAIL – получение и отправка электронной почты;

FTP – работа с файловым хранилищем.

В рамках эксперимента была поставлена задача собрать сбалансированный набор данных, содержащий примерно одинаковое количество записей для каждого класса. Для этой цели были отфильтрованы сессии для ряда категорий. Результаты проверки датасета на сбалансированность приведены в табл. 2.

Таблица 2. Распределение категорий и субкатегорий в датасете

fields.category	**ATTACK**	**USER_ACTION**
fields.sub_category		
ACCOUNT_MANIPULATION_(T1098)	125751	
CREDENTIAL_BRUTEFORCE_(T1110)	131530	
BITS_JOBS_(T1197)	124139	
SYSTEM_BINARY_PROXY_EXECUTION_(T1218)	102152	
OFFICE		136035
WEB		124682
MAIL		94057
FTP		83431
Всего	483572	438205

Анализ распределения категорий и субкатегорий в датасете, представленный в табл. 2, позволяет оценить степень сбалансированности набора данных, собранного в рамках эксперимента. Датасет включает две основные категории: ATTACK (атаки) и USER_ACTION (пользовательские действия), каждая из которых разделена на субкатегории. Общий объем данных составляет 921 777 уникальных записей (483 572 для ATTACK, что составляет 52,5% от общего объема и 438 205 для USER_ACTION – 47,5% от общего объема), что свидетельствует о достижении равновесия между классами. Незначительное преобладание ATTACK (на 5%) обусловлено особенностями сбора данных, поскольку атаки генерируют больше событий в журналах по сравнению с пользовательскими действиями.

Категория ATTACK включает четыре субкатегории, соответствующие техникам из матрицы MITRE ATT&CK:

ACCOUNT_MANIPULATION_(T1098): 125 751 запись (26% от ATTACK);

CREDENTIAL_BRUTEFORCE_(T1110): 131 530 записей (27,2% от ATTACK);

BITS_JOBS_(T1197): 124 139 записей (25,7% от ATTACK);

SYSTEM_BINARY_PROXY_EXECUTION_(T1218): 102 152 записи (21,1% от ATTACK).

Распределение субкатегорий ATTACK демонстрирует высокую степень сбалансированности: диапазон варьируется от 102 152 до 131 530 записей, с коэффициентом вариации около 9,5%. Это свидетельствует об успешной фильтрации данных для минимизации дисбаланса, что важно для предотвращения переобучения моделей на наиболее представленных классах.

Наибольшее количество записей приходится на CREDENTIAL_BRUTEFORCE_(T1110), что отражает частоту реализации этой техники в реальных атаках, в то время как SYSTEM_BINARY_PROXY_EXECUTION_(T1218) представлено наименьшим объемом ввиду более специфических условий ее применения. Такая вариативность подчеркивает необходимость тщательной фильтрации сессий для поддержания баланса при формировании датасетов, обеспечивая равные возможности для моделирования различных векторов угроз.

Категория USER_ACTION охватывает четыре субкатегории, отражающие типичные пользовательские активности в операционной системе:

OFFICE: 136 035 записей (31% от USER_ACTION);

WEB: 124 682 записи (28,4% от USER_ACTION);

MAIL: 94 057 записей (21,5% от USER_ACTION);

FTP: 83 431 запись (19,1% от USER_ACTION).

Распределение субкатегорий USER_ACTION также демонстрирует стремление к максимальной сбалансированности: диапазон варьируется от 83 431 до 136 035 записей, с коэффициентом вариации около 18%. Это указывает на менее идеальный баланс, что может быть обусловлено естественными различиями в частоте использования приложений в реальных сценариях – например, OFFICE и WEB, вероятно, генерируют больше событий из-за повседневного характера работы с офисными программами и веб-браузером, в то время как MAIL и FTP могут быть менее интенсивными или зависеть от специфических задач. Наибольший объем приходится на OFFICE, что отражает преобладание офисных задач в корпоративных средах, а наименьший – на FTP, возможно, из-за ограниченного использования файловых хранилищ в сравнении с другими действиями.

В целом, датасет достигает близкой к равномерной представленности классов на уровне категорий (52,5% ATTACK и 47,5% USER_ACTION) и субкатегорий, с коэффициентами вариации 9,5% для ATTACK и 18% для USER_ACTION. Это минимизирует смещение в моделях машинного обучения, обеспечивая репрезентативность для различения вредоносной активности и легитимных действий на основе EventID в системных журналах. Стремление к сбалансированности датасета обусловлена повышением точности обнаружения угроз в реальных системах. Такие распределения подтверждают эффективность формальной модели, где каждая активность оставляет уникальный след, и способствуют надежному анализу событий.

На рис. 3–5 наглядно представлено распределение 10 часто встречающихся EventID для каждого вида активности в журналах Security, System и Application.

Для наглядности распределения событий по всем категориям удобно использовать тепловую карту, как это представлено на рис. 6.

Уникальные наборы EventID для каждого вида компьютерной атаки и пользовательского действия представлены в табл. 3.

В ходе эксперимента установлено, что каждому виду компьютерной атаки или пользовательского действия соответствует уникальный набор значений EventID из системных журналов Security, System и Application операционной системы Windows. Данный факт позволяет использовать полученный датасет для обучения моделей машинного обучения и применения в современных интеллектуальных средствах защиты информации.

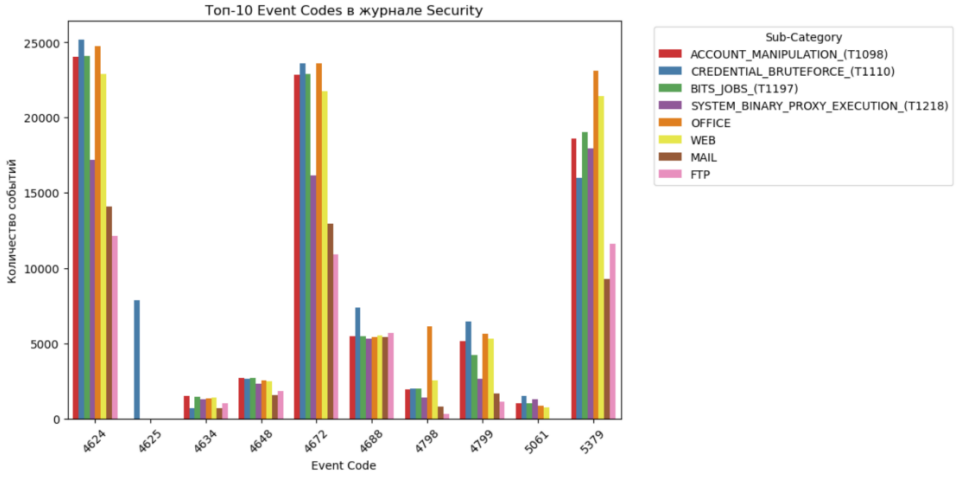

Рис. 3. Распределение EventID в журнале Security

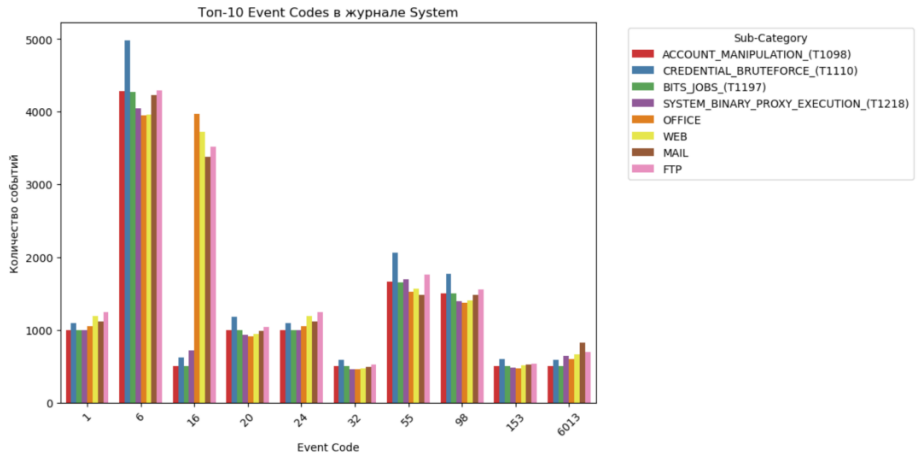

Рис. 4. Распределение EventID в журнале System

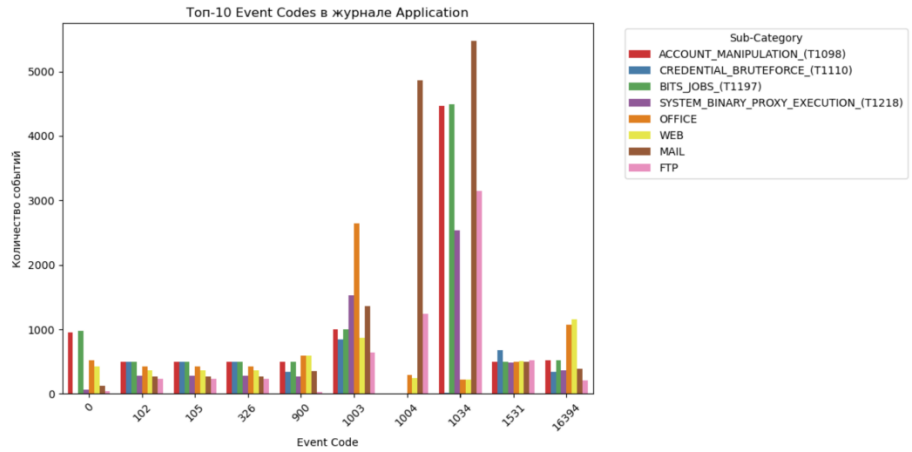

Рис. 5. Распределение EventID в журнале Application

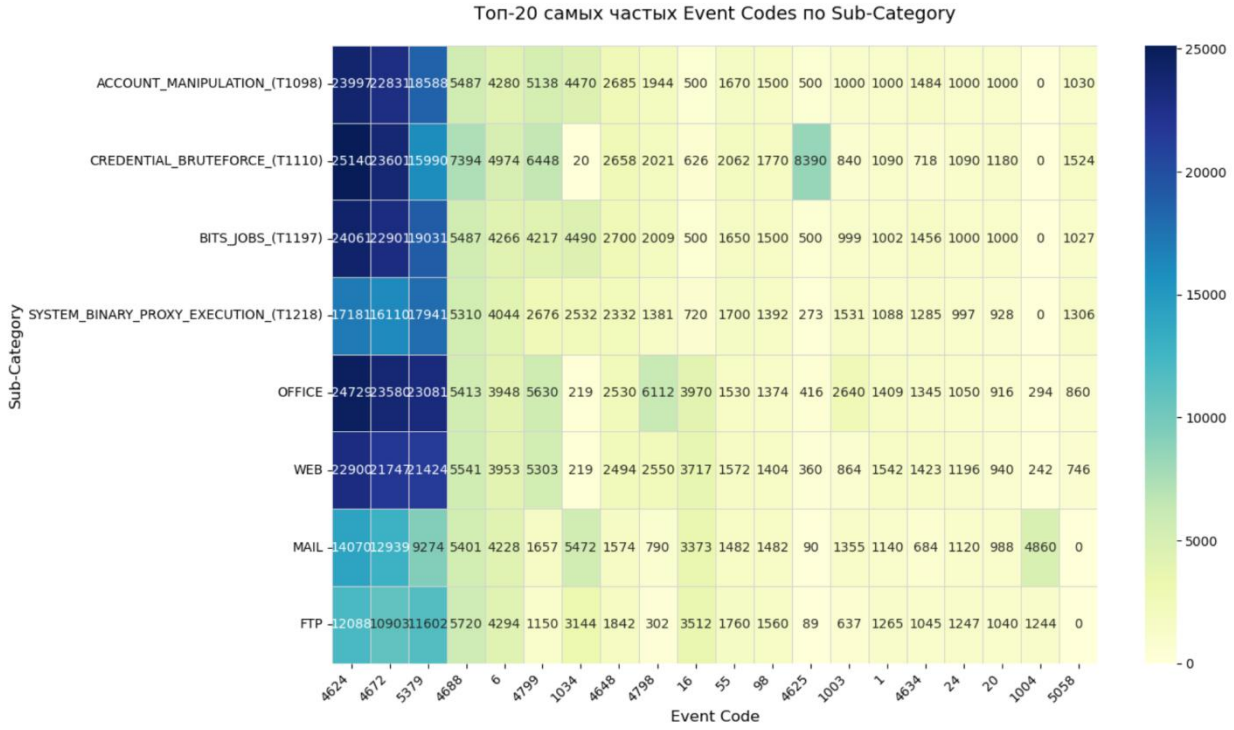

Рис. 6. Распределение EventID по всем категориям

Таблица 3. Наборы EventID для каждой категории

fields.sub_category	**event.code**
ACCOUNT_ MANIPULATION_ (T1098)	0, 1, 6, 12, 14, 16, 18, 20, 24, 25, 27, 32, 55, 98, 102, 105, 153, 172, 326, 900, 902, 903, 1000, 1001, 1003, 1033, 1034, 1066, 1531, 4608, 4616, 4624, 4625, 4634, 4648, 4672, 4688, 4696, 4738, 4781, 4798, 4799, 4826, 4902, 5024, 5033, 5058, 5059, 5061, 5379, 5382, 5615, 5617, 6005, 6009, 6013, 7001, 7026, 9027, 10010, 10016, 16384, 16394, 16962, 16977, 16983, 50036, 50103, 51046
CREDENTIAL_ BRUTEFORCE_(T1110)	1, 6, 12, 14, 15, 16, 18, 20, 24, 25, 27, 32, 41, 55, 98, 102, 105, 129, 153, 172, 256, 326, 508, 900, 902, 903, 1000, 1001, 1003, 1014, 1033, 1034, 1066, 1101, 1130, 1531, 4608, 4616, 4624, 4625, 4634, 4648, 4672, 4688, 4696, 4798, 4799, 4826, 4902, 5024, 5033, 5058, 5059, 5061, 5379, 5382, 5615, 5617, 6005, 6008, 6009, 6013, 7001, 7026, 7045, 8197, 8230, 9027, 10010, 10016, 12288, 12289, 16384, 16394, 16962, 16977, 16983, 50036, 50103, 51046
BITS_JOBS_(T1197)	0, 1, 6, 12, 14, 16, 18, 20, 24, 25, 27, 32, 55, 98, 102, 105, 129, 153, 172, 326, 900, 902, 903, 1000, 1001, 1003, 1033, 1034, 1066, 1531, 4608, 4616, 4624, 4625, 4634, 4648, 4672, 4688, 4696, 4798, 4799, 4826, 4902, 5024, 5033, 5058, 5059, 5061, 5379, 5382, 5615, 5617, 6005, 6009, 6013, 7001, 7026, 9027, 10010, 10016, 16384, 16394, 16962, 16977, 16983, 50036, 50103, 51046
SYSTEM_BINARY_PROXY_ EXECUTION_(T1218)	0, 1, 6, 12, 14, 15, 16, 18, 20, 24, 25, 27, 32, 34, 35, 37, 41, 52, 55, 98, 102, 105, 129, 153, 172, 256, 300, 301, 302, 326, 507, 508, 900, 902, 903, 1000, 1001, 1003, 1014, 1024, 1033, 1034, 1040, 1066, 1101, 1130, 1531, 4199, 4608, 4616, 4624, 4625, 4634, 4648, 4672, 4688, 4696, 4798, 4799, 4826, 4902, 5024, 5033, 5058, 5059, 5061, 5379, 5382, 5611, 5615, 5617, 6005, 6008, 6009, 6013, 7000, 7001, 7009, 7026, 7040, 7045, 8197, 8230, 9027, 10000, 10001, 10010, 10016, 10400, 12288, 12289, 16384, 16394, 16962, 16977, 16983, 50036, 50103, 51046
OFFICE	0, 1, 6, 12, 14, 15, 16, 18, 20, 24, 25, 27, 29, 32, 34, 35, 37, 41, 52, 55, 98, 102, 103, 105, 129, 153, 172, 256, 300, 301, 302, 326, 507, 508, 900, 902, 903, 1001, 1003, 1004, 1013, 1014, 1033, 1034, 1040, 1042, 1066, 1101, 1130, 1531, 4199, 4608, 4616, 4624, 4625, 4634, 4648, 4672, 4688, 4696, 4798, 4799, 4826, 4902, 5024, 5033, 5058, 5059, 5061, 5379, 5382, 5615, 5617, 6005, 6008, 6009, 6013, 7000, 7001, 7009, 7026, 7040, 7045, 8197, 8198, 8224, 8230, 9027, 10000, 10001, 10010, 10016, 10400, 11707, 12288, 12289, 16384, 16394, 16962, 16977, 16983, 50036, 50103, 51046
WEB	0, 1, 6, 12, 14, 15, 16, 18, 20, 24, 25, 27, 32, 34, 35, 37, 41, 43, 44, 55, 98, 102, 103, 105, 129, 134, 153, 172, 256, 300, 301, 302, 326, 508, 900, 902, 903, 1000, 1001, 1002, 1003, 1004, 1013, 1033, 1034, 1040, 1042, 1066, 1101, 1130, 1531, 4000, 4100, 4608, 4616, 4624, 4625, 4634, 4648, 4672, 4688, 4696, 4797, 4798, 4799, 4826, 4902, 5024, 5033, 5058, 5059, 5061, 5379, 5382, 5615, 5617, 6005, 6008, 6009, 6013, 7001, 7023, 7026, 7040, 7045, 8197, 8198, 8224, 8230, 9027, 10000, 10001, 10010, 10016, 11707, 12288, 12289, 16384, 16394, 16962, 16977, 16983, 50036, 50103, 51046
MAIL	0, 1, 6, 12, 14, 15, 16, 18, 20, 24, 25, 27, 32, 34, 37, 41, 55, 98, 102, 103, 105, 129, 153, 172, 256, 300, 301, 302, 326, 508, 510, 900, 902, 903, 1003, 1004, 1013, 1033, 1034, 1035, 1040, 1042, 1066, 1101, 1531, 4608, 4616, 4624, 4625, 4634, 4648, 4672, 4688, 4696, 4798, 4799, 4826, 4902, 5024, 5033, 5379, 5615, 5617, 6005, 6008, 6009, 6013, 7000, 7001, 7009, 7026, 7040, 7045, 8197, 8230, 10000, 10001, 10010, 10016, 11728, 12288, 12289, 16384, 16394, 16962, 16977, 16983, 50036, 50103, 51046
FTP	0, 1, 6, 12, 14, 15, 16, 18, 20, 24, 25, 27, 32, 34, 35, 37, 41, 55, 98, 102, 103, 105, 129, 153, 172, 256, 300, 301, 302, 326, 508, 900, 902, 903, 1003, 1004, 1013, 1033, 1034, 1035, 1040, 1042, 1066, 1101, 1531, 4608, 4616, 4624, 4625, 4634, 4648, 4672, 4688, 4696, 4798, 4799, 4826, 4902, 5024, 5033, 5379, 5382, 5615, 5617, 6005, 6008, 6009, 6013, 7001, 7026, 7040, 7045, 8197, 8230, 10000, 10001, 10010, 10016, 11728, 12288, 12289, 16384, 16394, 16962, 16977, 16983, 50036, 50103, 51046

Заключение

В работе проведён обзор современных методов выявления компьютерных атак, основанных на анализе лог-файлов и машинном обучении, что позволило выявить преимущества и ограничения существующих подходов.

Рассмотрены существующие подходы к моделированию атак и формированию наборов данных, такие как симуляции с использованием фреймворков типа MITRE ATT&CK и инструментов для генерации синтетических данных, что обеспечило основу для разработки собственного решения.

Разработан алгоритм формирования базы данных из записей системных журналов, учитывающий различные сценарии атак (включая техники MITRE ATT&CK) и пользовательской активности, с интеграцией валидации скриптов и метаданных для повышения достоверности.

Спроектирована и реализована виртуальная сетевая инфраструктура, обеспечивающая автоматизированный сбор и обработку системных журналов.

Проведены эксперименты по формированию датасета объёмом 921 777 записей, с оценкой сбалансированности (52,5% ATTACK и 47,5% USER_ACTION) и потенциала для обучения моделей машинного обучения, подтвердившего репрезентативность и применимость данных.

Цель исследования достигнута – авторами разработан алгоритм, способствующий автоматизации сбора данных, более полному покрытию тактик MITRE ATT&CK и повышению достоверности, а также сформирован сбалансированный размеченный датасет. Полученный набор данных необходим для обучения и тестирования моделей машинного обучения, предназначенных для автоматического выявления киберугроз в системных логах, что позволяет специалистам в области информационной безопасности разрабатывать более точные системы обнаружения атак, снижая ложные срабатывания и улучшая защиту от реальных угроз, а экспертам в области искусственного интеллекта совершенствовать алгоритмы обработки больших данных и анализа аномалий.

СПИСОК ЛИТЕРАТУРЫ:

1. Qadeer M.A. et al. Network traffic analysis and intrusion detection using packet sniffer. Second International Conference on Communication Software and Networks. IEEE, 2010, p. 313–317. DOI: http://dx.doi.org/10.1109/ICCSN.2010.104.
2. He S. et al. Experience report: System log analysis for anomaly detection. IEEE 27th international symposium on software reliability engineering (ISSRE). IEEE, 2016, p. 207–218. DOI: http://dx.doi.org/10.1109/ISSRE.2016.21.
3. Nehinbe J.O. Log Analyzer for Network Forensics and Incident Reporting. International Conference on Intelligent Systems, Modelling and Simulation. IEEE, 2010, p. 356–361. DOI: http://dx.doi.org/10.1109/ISMS.2010.71.
4. Грибачёв А.С., Кальщиков В.В., Ручай А.Н. Методы, алгоритмы и базы данных обнаружения компьютерных инцидентов. Вестник УрФО. Безопасность в информационной сфере. 2024, т. 1, № 51, с. 45–52. URL: http://info-secur.ru/index.php/ojs/article/view/441/399 (дата обращения: 18.09.2025).
5. Павлычев А.В., Стародубов М.И. Формирование набора данных в задаче обнаружения компьютерных атак с использованием методов машинного обучения.Современное образование: интеграция образования, науки, бизнеса и власти. Трансформация образования, науки и производства-основа технологического прорыва. 2023, с. 161–166. URL: https://elibrary.ru/item.asp?id=51907642 (дата обращения: 18.09.2025).
6. Lippmann R.P. et al. Evaluating intrusion detection systems: The 1998 DARPA off-line intrusion detection evaluation. Proceedings DARPA Information Survivability Conference and Exposition. DISCEX'00. IEEE, 2000, v. 2, p. 12–26. DOI: http://dx.doi.org/10.1109/DISCEX.2000.821506.
7. Brown C. et al. Analysis of the 1999 darpa/lincoln laboratory ids evaluation data with netadhict. IEEE Symposium on Computational Intelligence for Security and Defense Applications. IEEE, 2009, p. 1–7. DOI: http://dx.doi.org/10.1007/978-3-540-45248-5_13.

8. Bolon-Canedo V., Sanchez-Marono N., Alonso-Betanzos A. Feature selection and classification in multiple class datasets: An application to KDD Cup 99 dataset. Expert Systems with Applications. 2011, p. 38, no. 5, p. 5947–5957. DOI: http://dx.doi.org/10.1016/j.eswa.2010.11.028.
9. Tavallaee M. et al. A detailed analysis of the KDD CUP 99 data set. IEEE symposium on computational intelligence for security and defense applications. IEEE, 2009, p. 1–6. DOI: http://dx.doi.org/10.1109/CISDA.2009.5356528.
10. Dhanabal L., Shantharajah S.P. A study on NSL-KDD dataset for intrusion detection system based on classification algorithms. International journal of advanced research in computer and communication engineering. 2015, v. 4, no. 6, p. 446–452. DOI: http://dx.doi.org/10.17148/IJARCCE.2015.4696.
11. Ghurab M. et al. A detailed analysis of benchmark datasets for network intrusion detection system. Asian Journal of Research in Computer Science. 2021, v. 7, no. 4, p. 14–33. URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3834787 (дата обращения: 18.09.2025).
12. Carvalho D.V., Pereira E.M., Cardoso J.S. Machine learning interpretability: A survey on methods and metrics. Electronics. 2019, v. 8, no. 8, p. 832. DOI: http://dx.doi.org/10.3390/electronics8080832.
13. Amrollahi M. et al. Enhancing network security via machine learning: opportunities and challenges. Handbook of big data privacy. 2020, p. 165–189. DOI: http://dx.doi.org/10.1007/978-3-030-38557-6_8.
14. Banerjee J. et al. Impact of machine learning in various network security applications. 3rd International conference on computing methodologies and communication (ICCMC). IEEE, 2019, p. 276–281. DOI: http://dx.doi.org/10.1109/ICCMC.2019.8819811.
15. Leblanc S.P. et al. An overview of cyber attack and computer network operations simulation. Proceedings of the 2011 Military Modeling & Simulation Symposium. 2011, p. 92–100. DOI: http://dx.doi.org/10.5555/2048558.2048572.
16. Strom B.E. et al. Mitre att&ck: Design and philosophy. Technical report. The MITRE Corporation, 2018. URL: https://www.mitre.org/sites/default/files/2021-11/prs-19-01075-28-mitre-attack-design-and-philosophy.pdf (дата обращения: 18.09.2025).
17. Dwyer J., Truta T. M. Finding anomalies in windows event logs using standard deviation. 9th IEEE International Conference on Collaborative Computing: Networking, Applications and Worksharing. 2013, p. 563–570. DOI: http://dx.doi.org/10.4108/icst.collaboratecom.2013.254136.
18. Smiliotopoulos C., Kambourakis G., Barbatsalou K. On the detection of lateral movement through supervised machine learning and an open-source tool to create turnkey datasets from Sysmon logs. International Journal of Information Security. 2023, v. 22, no. 6, p. 1893–1919. DOI: http://dx.doi.org/10.1007/s10207-023-00725-8.
19. Smiliotopoulos C., Kambourakis G., Kolias C. Detecting lateral movement: A systematic survey. Heliyon. 2024, v. 10, no. 4. DOI: http://dx.doi.org/10.1016/j.heliyon.2024.e26317.
20. Smiliotopoulos C., Barmpatsalou K., Kambourakis G. Revisiting the detection of lateral movement through Sysmon. Applied Sciences. 2022, v. 12, no. 15, p. 7746. DOI: http://dx.doi.org/10.3390/app12157746.
21. Rahal K., Riahi A., Debatty T. Dataset of APT Persistence Techniques on Windows Platforms Mapped to the MITRE ATT&CK Framework. 28th Conference on Innovation in Clouds, Internet and Networks (ICIN). IEEE, 2025, p. 17–24. DOI: http://dx.doi.org/10.1109/ICIN64016.2025.10943025.
22. Husselman L. Anomaly Detection with Windows Event Logs: A comparative study between traditional and ML based approaches. University of Zurich, 2024. DOI: http://dx.doi.org/10.5167/uzh-264648.

REFERENCES:

[1] Qadeer M.A. et al. Network traffic analysis and intrusion detection using packet sniffer. Second International Conference on Communication Software and Networks. IEEE, 2010, p. 313–317. DOI: http://dx.doi.org/10.1109/ICCSN.2010.104.
[2] He S. et al. Experience report: System log analysis for anomaly detection. IEEE 27th international symposium on software reliability engineering (ISSRE). IEEE, 2016, p. 207–218. DOI: http://dx.doi.org/10.1109/ISSRE.2016.21.
[3] Nehinbe J.O. Log Analyzer for Network Forensics and Incident Reporting. International Conference on Intelligent Systems, Modelling and Simulation. IEEE, 2010, p. 356–361. DOI: http://dx.doi.org/10.1109/ISMS.2010.71.
[4] Gribachev A.S., Kalshchikov V.V., Ruchai A.N. Methods, algorithms and databases for detecting computer incidents. Bulletin of the Ural Federal District. Information security. 2024, v. 1, no. 51, p. 45–52. URL: http://info-secur.ru/index.php/ojs/article/view/441/399 (accessed: 18.09.2025) (in Russian).
[5] Pavlychev A.V., Starodubov M.I. Data set formation in the task of detecting computer attacks using machine learning methods. Modern education: integration of education, science, business and government. The transformation of education, science and production is the basis for a technological breakthrough. 2023, p. 161–166. URL: https://elibrary.ru/item.asp?id=51907642 (accessed: 18.09.2025) (in Russian).

[6] Lippmann R.P. et al. Evaluating intrusion detection systems: The 1998 DARPA off-line intrusion detection evaluation. Proceedings DARPA Information Survivability Conference and Exposition. DISCEX'00. IEEE, 2000, v. 2, p. 12–26. DOI: http://dx.doi.org/10.1109/DISCEX.2000.821506.
[7] Brown C. et al. Analysis of the 1999 darpa/lincoln laboratory ids evaluation data with netadhict. IEEE Symposium on Computational Intelligence for Security and Defense Applications. IEEE, 2009, p. 1–7. DOI: http://dx.doi.org/10.1007/978-3-540-45248-5_13.
[8] Bolon-Canedo V., Sanchez-Marono N., Alonso-Betanzos A. Feature selection and classification in multiple class datasets: An application to KDD Cup 99 dataset. Expert Systems with Applications. 2011, p. 38, no. 5, p. 5947–5957. DOI: http://dx.doi.org/10.1016/j.eswa.2010.11.028.
[9] Tavallaee M. et al. A detailed analysis of the KDD CUP 99 data set. IEEE symposium on computational intelligence for security and defense applications. IEEE, 2009, p. 1–6. DOI: http://dx.doi.org/10.1109/CISDA.2009.5356528.
[10] Dhanabal L., Shantharajah S.P. A study on NSL-KDD dataset for intrusion detection system based on classification algorithms. International journal of advanced research in computer and communication engineering. 2015, v. 4, no. 6, p. 446–452. DOI: http://dx.doi.org/10.17148/IJARCCE.2015.4696.
[11] Ghurab M. et al. A detailed analysis of benchmark datasets for network intrusion detection system. Asian Journal of Research in Computer Science. 2021, v. 7, no. 4, p. 14–33. URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3834787 (дата обращения: 18.09.2025).
[12] Carvalho D.V., Pereira E.M., Cardoso J.S. Machine learning interpretability: A survey on methods and metrics. Electronics. 2019, v. 8, no. 8, p. 832. DOI: http://dx.doi.org/10.3390/electronics8080832.
[13] Amrollahi M. et al. Enhancing network security via machine learning: opportunities and challenges. Handbook of big data privacy. 2020, p. 165–189. DOI: http://dx.doi.org/10.1007/978-3-030-38557-6_8.
[14] Banerjee J. et al. Impact of machine learning in various network security applications. 3rd International conference on computing methodologies and communication (ICCMC). IEEE, 2019, p. 276–281. DOI: http://dx.doi.org/10.1109/ICCMC.2019.8819811.
[15] Leblanc S.P. et al. An overview of cyber attack and computer network operations simulation. Proceedings of the 2011 Military Modeling & Simulation Symposium. 2011, p. 92–100. DOI: http://dx.doi.org/10.5555/2048558.2048572.
[16] Strom B.E. et al. Mitre att&ck: Design and philosophy. Technical report. The MITRE Corporation, 2018. URL: https://www.mitre.org/sites/default/files/2021-11/prs-19-01075-28-mitre-attack-design-and-philosophy.pdf (дата обращения: 18.09.2025).
[17] Dwyer J., Truta T. M. Finding anomalies in windows event logs using standard deviation. 9th IEEE International Conference on Collaborative Computing: Networking, Applications and Worksharing. 2013, p. 563–570. DOI: http://dx.doi.org/10.4108/icst.collaboratecom.2013.254136.
[18] Smiliotopoulos C., Kambourakis G., Barbatsalou K. On the detection of lateral movement through supervised machine learning and an open-source tool to create turnkey datasets from Sysmon logs. International Journal of Information Security. 2023, v. 22, no. 6, p. 1893–1919. DOI: http://dx.doi.org/10.1007/s10207-023-00725-8.
[19] Smiliotopoulos C., Kambourakis G., Kolias C. Detecting lateral movement: A systematic survey. Heliyon. 2024, v. 10, no. 4. DOI: http://dx.doi.org/10.1016/j.heliyon.2024.e26317.
[20] Smiliotopoulos C., Barmpatsalou K., Kambourakis G. Revisiting the detection of lateral movement through Sysmon. Applied Sciences. 2022, v. 12, no. 15, p. 7746. DOI: http://dx.doi.org/10.3390/app12157746.
[21] Rahal K., Riahi A., Debatty T. Dataset of APT Persistence Techniques on Windows Platforms Mapped to the MITRE ATT&CK Framework. 28th Conference on Innovation in Clouds, Internet and Networks (ICIN). IEEE, 2025, p. 17–24. DOI: http://dx.doi.org/10.1109/ICIN64016.2025.10943025.
[22] Husselman L. Anomaly Detection with Windows Event Logs: A comparative study between traditional and ML based approaches. University of Zurich, 2024. DOI: http://dx.doi.org/10.5167/uzh-264648.

Статья поступила в редакцию 07.09.2025; одобрена после рецензирования 07.10.2025;
принята к публикации 07.11.2025
The article was submitted 07.09.2025; approved after reviewing 07.10.2025;
accepted for publication 07.11.2025