ISSN 2305-994X

МАТЕРИАЛЫ
35-Й НАУЧНО-ТЕХНИЧЕСКОЙ ВСЕРОССИЙСКОЙ КОНФЕРЕНЦИИ

МЕТОДЫ И ТЕХНИЧЕСКИЕ СРЕДСТВА ОБЕСПЕЧЕНИЯ БЕЗОПАСНОСТИ ИНФОРМАЦИИ

23–26 ИЮНЯ 2026 ГОДА

Санкт-Петербург

2026

Методы и технические средства обеспечения безопасности информации:
Материалы 35-й научно-технической всероссийской конференции 23–26 июня 2026 года. СПб: ПОЛИТЕХ-ПРЕСС, 2026. 263 с.

Приводятся тезисы докладов, отражающие теоретические и практические проблемы обеспечения безопасности информационных технологий, а также вопросы подготовки и переподготовки специалистов в этом направлении.

Предназначаются для преподавателей, научных сотрудников и инженерно-технических работников, занимающихся проблемами информационной безопасности. Ответственный за выпуск – доктор технических наук, профессор, чл.-кор. РАН Д.П. Зегжда.

Сборник печатается без редакторских правок.

Ответственность за содержание тезисов возлагается на авторов.

ISSN 2305-994X

1. ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ В ЗАДАЧАХ КИБЕРБЕЗОПАСНОСТИ: ТЕОРИЯ И ПРАКТИКА

Абитов Р.А., Дахнович А.Д., Москвин Д.А.
ООО «НеоБИТ», Санкт-Петербург

ИССЛЕДОВАНИЕ МЕТОДОВ АВТОМАТИЗИРОВАННОГО ИЗВЛЕЧЕНИЯ ДАННЫХ ИЗ ВЕБ-РЕСУРСОВ С ДИНАМИЧЕСКИ ИЗМЕНЯЕМЫМ КОНТЕНТОМ С ПОМОЩЬЮ ИИ-АГЕНТОВ

В работе рассматривается задача автоматизированного извлечения данных из веб-ресурсов с динамически изменяемым контентом с применением ИИ-агентов. Актуальность исследования обусловлена ростом потребности в надёжном и масштабируемом сборе данных для обучения интеллектуальных систем, информационно-аналитических задач, статистических исследований, а также финансового и экономического прогнозирования.

Основная проблема заключается в выборе оптимальной нейросетевой модели и конфигурации агентской системы по совокупности критериев: стоимость, быстродействие, асинхронность обработки, отказоустойчивость и качество извлекаемых данных. Цель исследования состоит в определении наилучшего метода извлечения данных из веб-ресурсов с динамически изменяемым контентом с использованием ИИ-агентов. Объект исследования — веб-ресурсы с динамически изменяемым контентом. Предмет исследования — методы автоматизированного извлечения данных из таких веб-ресурсов на базе ИИ-агентов.

Предлагаемый метод разделяет дорогостоящий этап анализа веб-ресурса с помощью ИИ и экономичный этап массового сбора. Метод включает изолированный браузерный контур на базе Chromium, ориентированный на устойчивую автоматизацию доступа с помощью средств сокрытия процессов автоматизации, и подключение ИИ-агента к инструментам управления браузеров с помощью Model Context Protocol (MCP) [1]. На этапе конфигурирования используется агент OpenClaw: он анализирует целевой ресурс (навигация, стабилизация и унификация DOM, обнаружение необходимых данных), формализует схему полей и синтезирует набор статических средств сбора — сценарии и инструменты (tools), после чего регулярный сбор данных выполняется без обращения к LLM на каждую единицу извлекаемой информации, что снижает совокупную стоимость инференса и повышает пропускную способность. Метод рассматривает задачи генерации переносимых правил сбора с опорой на языковую модель и структуру HTML [2]; перевод интерактивного извлечения в исполнимые процедуры для масштабирования по объёму данных [3]; для методологии оценки автономных веб-агентов в реалистичной среде целесообразно опираться на согласованные постановки измерения успеха и типовые ограничения современных систем оценки ИИ агентов для управления браузерами [4]; для

модульной организации многошагового контура «поиск — извлечение — агрегация» — используются агентные схемы веб-агрегации [5].

В рамках исследования был выполнен сравнительный анализ языковых моделей и архитектур агентских систем, классификация сайтов по категориям данных и разработка методических шаблонов, инструментов сбора под каждый класс (списки и таблицы, карточные ленты, фильтры и пагинация, подгрузка контента по событиям и др.). Произведена оценка качества, задержек и стоимости на единицу успешно извлечённых данных, а также анализ отказоустойчивости при изменении вёрстки. Ожидаемым результатом является разработка методики по выбору модели, настройке агентской системы, обеспечивающих экономию средств, скорость сбора, качество и объём данных при снижении трудозатрат на сопровождение.

Ключевые слова: ИИ-агенты, извлечение данных, веб-скрапинг, динамический контент, нейронные сети, асинхронная обработка, отказоустойчивость, качество данных, автоматизация, Model Context Protocol, Chromium, OpenClaw.

Список источников

1. Introducing the Model Context Protocol // Anthropic – https://www.anthropic.com/news/model-context-protocol – 2024 (дата обращения 16.05.2026).
2. Huang W. et al. AutoScraper: A Progressive Understanding Web Agent for Web Scraper Generation // arXiv.org – https://arxiv.org/abs/2404.12753 – 2024 (дата обращения 16.05.2026).
3. Дахнович А. Д., Богина В. М., Макеева А. А. Применение больших языковых моделей в задаче прогнозирования событий // Проблемы информационной безопасности. Компьютерные системы. 2025. № S2. С. 58–68. (дата обращения 16.05.2026).
4. Zhou S. et al. WebArena: A Realistic Web Environment for Building Autonomous Agents // arXiv.org – https://arxiv.org/abs/2307.13854 – 2023 (дата обращения 16.05.2026).
5. Reddy R. G. et al. Infogent: An Agent-Based Framework for Web Information Aggregation // arXiv.org – https://arxiv.org/abs/2410.19054 – 2024 (дата обращения 16.05.2026).

Баранов А.П.
Научно-исследовательский институт системных исследований РАН, Москва

ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ И ЭКСПЕРТНЫЕ СИСТЕМЫ В БЕЗОПАСНОСТИ ИНФОРМАЦИИ

Результаты и оценки, полученные при применении компьютерных систем, всегда должны быть подвергнуты окончательной верификации квалифицированными специалистами.

1. Определения искусственного интеллекта и экспертных систем

Искусственный интеллект (ИИ) — это комплекс технологических решений, позволяющий имитировать когнитивные функции человека (включая самообучение и поиск решений без заранее заданного алгоритма) и получать при выполнении конкретных задач результаты, сопоставимые, как минимум, с результатами интеллектуальной деятельности человека. Данный комплекс включает в себя информационно-

коммуникационную инфраструктуру, программное обеспечение (в том числе с использованием методов машинного обучения), процессы и сервисы по обработке данных и поиску решений.

Экспертная система (ЭС) — это программный комплекс, который воспроизводит процесс решения проблемы человеком-экспертом. В отличие от обычных алгоритмических программ, экспертные системы оперируют не просто данными, а знаниями, представленными в символьной форме. Ключевое отличие экспертной системы от других программ состоит в способности объяснять ход своих рассуждений понятным для пользователя образом. Это не «чёрный ящик», а прозрачная система логических выводов.

2. Методологические основания: системный подход и философия познания

Философия как особая форма познания и система знаний об общих характеристиках, понятиях и принципах реальности является методологической основой любого научного направления. Любой логический вывод делается на основе применения логических правил и основывается на первичных определениях или постулатах. Системный подход предполагает представление жизненной или технической ситуации в виде множества взаимодействующих систем.

Взаимодействие (отражение) — воздействие одной системы на другую, одностороннее или обратное, либо на внешнюю среду, которую представляют, как другую систему. Воздействие является первичным понятием. Модель системы (М) — это упрощённое представление системы, отражающее наиболее существенные элементы, связи и процессы, важные для анализа, проектирования, управления или понимания поведения системы в определённом контексте.

3. Понятия безопасности системы и безопасности информации

Безопасность — это специфическая совокупность условий существования систем, в том числе человеческих, контролируемых ею и обеспечивающих: сохранение их целостности, устойчивое развитие (либо стагнацию или предотвращение деградации) и эффективную реализацию присущего им функционала, включая взаимодействие (отражение) между системами или изолированное существование в соответствующей среде.

Информация — это содержание взаимодействия систем или разнообразие самой системы (по Урсулу А. Д.). Информация существует только при взаимодействии (отображении) и только при её движении. Под движением понимается вся совокупность изменений: изменение информации, включая перемещение на материальных носителях или во времени, есть движение. Безопасность информации (БИ) — это совокупность условий существования информации, обеспечивающих сохранение её целостности и устойчивого движения в целях обеспечения эффективного взаимодействия систем. Следствием является обеспечение конфиденциальности, целостности и доступности (К, Ц, Д).

4. Интеллект системы, ИИ и ЭС

Естественный интеллект системы (ЕИС) — это способность отражающей системы к построению моделей действительности, синтезу новых, возможно ещё не существующих, а также оценка степени реализации присущего системе функционала в связи с параметрами модели. Знание — это модель, способы синтеза новых моделей (включая виртуальные), а также расчёт оценки возможности реализации функционала отражающей системы в условиях сформированной модели.

Моделирование — суть существования ЕИС. В составе оценок параметров модели, как правило, присутствует построение прогноза развития модели с верификацией предсказания. Принято различать конечные и континуальные модели. Модель ЕИС — это и есть ИИ или ЭС: системы с конечным множеством состояний, моделирующие ЕИ

методом реализации компьютерных алгоритмов. ЕИ легко оперирует бесконечным числом объектов, тогда как ИИ и ЭС работают только с конечным. Модель обязательно согласуется компетентным техническим специалистом как основа выводов.

5. ЕИС и ЭС в обеспечении безопасности

ЕИС моделирует условия своего существования с проекцией на них функционала системы. В этой модели ЕИ оценивает параметры безопасности, и если оценки негативны, вырабатывает новые модели. Набор угроз и условий их реализации — это действительность, которую отражает сформированная ЕИС-модель в области безопасности. Эта модель является частью знаний в данной области.

ЭС в безопасности — это моделирование ЕИС в части оценки условий безопасности, оценка параметров обеспечения безопасности. При несоблюдении условий безопасности осуществляется построение новых моделей действительности или изменение в системе. Конкретная задача ЭС в безопасности состоит в построении противодействия противнику, моделируемому в виде известного набора угроз и способов нападения. Модель угроз и условий для ИИ и ЭС всегда конечна и составляет часть базы знаний; другая часть — алгоритмы расчёта оценок безопасности и синтеза новых моделей. Специалист или регулятор определяет границы конечности.

6. Философская схема для системного анализа

В рамках системного анализа рассматривается взаимодействие отражающей и отражаемой систем. Информация понимается как содержание отражения (по Урсулу А. Д.). Безопасность системы — это состояние условий реализации её функционала. Безопасность информации — это условия безопасности движения содержания взаимодействия (отражения). Интеллект — отражение действительности методом моделирования, установление соответствия построенной модели предназначению системы.

ИИ и ЭС — конечные компьютерные модели ЕИС. ЭС в безопасности осуществляет анализ угроз реализации функционала системы, синтез новых моделей с последующей оценкой, выдачу предложений по изменению системы и переоценку угроз. Таким образом, ЭС — это компьютерная система на основе моделируемого представления знаний (угроз) и данных, обработки входных данных и вывода результатов.

7. Модели для ЭС анализа состояния информационной безопасности компьютерной системы

Модель компьютерной системы (КС) включает: структуру транспортной сети и программного обеспечения реализации сети связи; перечень серверов и систем хранения информации; идентификацию видов базового ПО (типов BIOS, виртуализаторов, ОС) и степени их доверенности; определение типов прикладного ПО на серверах и рабочих местах, а также способов их взаимодействия; правила допуска пользователей к частям информации.

Модель угроз базируется на перечне ФСТЭК и ФСБ для криптосистем, тем самым моделируя агрессивность внешней среды существования КС. Модель внутренних угроз формируется из конкретных условий применения КС, коллективов пользователей и обслуживающего персонала. Особую роль сегодня играет оперативно-техническая составляющая нападения на КС, которая пока мало прописана регуляторами. Морально-психологическое воздействие в сочетании с угрозами применения физических и материальных деяний существенно корректирует модель КС.

8. Представление знаний ИИ в информационной, компьютерной и криптографической безопасности

ИИ может реализовывать все три функции ЕИС. «Генеративный» ИИ способен генерировать новые знания по заложенному алгоритму. «Слабый» ИИ работает с входными

данными только на ранее введённой базе знаний. Обучение — это способ пополнения базы знаний; если после обучения база знаний в процессе применения не меняется, речь идёт о «Слабом» ИИ.

Примеры «Слабого» ИИ: набор угроз ИБ от ФСТЭК; набор угроз компьютерной безопасности как часть предыдущего; набор угроз криптографической безопасности, состоящий из моделей КС1–КА от ФСБ. Угрозы и способы их реализации являются базами знаний. Например, знания в криптографической безопасности — это перечень методов криптоанализа, модель шифра и угрозы из КС1–КА. «Слабый» ИИ использует фиксированную базу угроз и способов их реализации при функционировании модели защищаемой системы.

Самообучение и адаптация — суть «Генеративного» ИИ, который вырабатывает новые знания: угрозы и, возможно, модели — по заданному квалифицированным разработчиком алгоритму. «Слабый» ИИ при оценке состояния компьютерной безопасности предполагает: формализованное представление и структуризацию как системы угроз, так и защищаемой компьютерной системы (ЗКС) в связи с телекоммуникациями; формализацию привязки угроз к конкретным блокам или функциям; описание аппаратной части ЗКС как формальной совокупности блоков. Нейронная сеть или лингвистическая модель с предварительным обучением являются типичными представителями «Слабого» ИИ.

Для оценки доступности также требуется подобная схема. Вопрос о том, где найти модель угроз доступности, остаётся открытым: частичный ответ может содержаться в требованиях к безопасности критической информационной инфраструктуры (КИИ) и в следующих стандартах: ГОСТ Р 55388-2012, ГОСТ Р 55387-2012, ГОСТ Р 53731-2009, ГОСТ Р 53728-2009, ГОСТ Р 53724-2009.

Для «Генеративного» ИИ в компьютерной безопасности ключевым является вопрос о генерации новых знаний. Этот метод должен базироваться на способе представления знаний и учёте входных данных, включая представление ЗКС в виде блоков для возможности их реконфигурирования. Предлагаемый подход предусматривает расчленение угроз и методов их реализации на фрагменты, разработку интерфейсов соединения различных фрагментов и разработку метода генерации новых угроз на основе фрагментов базовой системы. Потенциальная опасность состоит в генерации фантастических, нереальных угроз при неточности выработки критериев непротиворечивости цепочки фрагментов. Поэтому необходима оценка реальности новых вариантов угроз экспертами на тестовых примерах.

9. Программные платформы для разработки ИИ в России

В настоящее время в России используются два зарубежных программных продукта из области фундаментальных библиотек для разработки ИИ-решений: PyTorch (Meta/Facebook) и TensorFlow (Google). Также применяются Orange (Франция, Google) в версии от Любляны (Словения) и SMILE.Cloud от ИТМО на AutoVL. Все эти продукты предназначены для построения многослойных нейронных сетей различной графовой конфигурации с функциями обратной и градиентной связи. При этом база знаний хранится исключительно в значениях весовых коэффициентов.

Открытые тексты библиотек ПО для нейронных сетей написаны на языках высокого уровня (Python, Java), которые представляют собой библиотеки функций, не сертифицированных по отсутствию недекларированных возможностей (НДВ). Математически верифицировать результат работы таких программ возможно лишь в простейших случаях: уже двухслойные сети без обратных связей не поддаются математическому анализу. Принцип верификации состоит в следующем: построение

математической модели данных, расчёт результата работы, сравнение теоретического результата с расчётным по программе.

10. Информационная безопасность ИИ: доверенный ИИ

Безопасность ИИ определяется как информационная безопасность компьютерной, информационно-аналитической системы, называемой ИИ (то есть обеспечение К, Ц, Д). Особенностью является то, что требования регуляторов ИБ на подсистемы — базу знаний, систему экспертной оценки, тестовые системы и обрабатываемые данные — в настоящее время отсутствуют. Оценка опасности и надёжности экспертной системы может быть произведена только в связи с отраслью применения.

Серьёзную угрозу представляет возможность преднамеренного переориентирования «Генеративного» ИИ на заведомо ложное поведение в процессе обучения на основе оценок его ответов. Экспериментально установлено, что ChatGPT дообучается на ответах пользователей, чья природа по существу неизвестна. Возникает принципиальный вопрос: как отличить правильное понятие от намеренно введённого корректно ложного? Не доверенный ИИ становится платформой для информационных диверсий и распространения ложных знаний и представлений.

11. Выводы и предложения

Все выводы и результаты работы ИИ и ЭС должны быть подвергнуты анализу компетентных экспертов — в первую очередь на предмет отсутствия логических ошибок — причём специалистов, не участвовавших в разработке данного ИИ или ЭС.

Рассматриваемые системы желательно представить регуляторам как наиболее компетентным экспертам в области компьютерной безопасности. Интеллектуальные системы ИИ или ЭС для оценки состояния информационной безопасности могут быть построены для различных систем: промышленных, производственных (как с АСУ ТП, так и без), общественных объединений и финансово-экономических структур.

Системы ИИ или ЭС являются сложной и высокозатратной разработкой. По мнению автора, в области компьютерной или криптографической безопасности подобное направление под силу реализовать только по государственному заказу со стороны регуляторов, если они сочтут такую разработку целесообразной.

Безбородов П.Д., Полтавцева М.А.
Санкт-Петербургский политехнический университет Петра Великого, Санкт-Петербург

ПРОБЛЕМА ОБЕСПЕЧЕНИЯ КОНФИДЕНЦИАЛЬНОСТИ МОДЕЛЕЙ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА В ФЕДЕРАТИВНОМ ОБУЧЕНИИ

Технологии искусственного интеллекта стремительно развиваются, в том числе и федеративное обучение, призванное вывести качество моделей искусственного интеллекта на новый уровень. Для федеративного обучения в настоящее время не представлены решения, способствующие полному обеспечению конфиденциальности моделей и, как следствие, данных обучающей выборки. В частности, остается без рассмотрения ситуация, в которой по отношению к одному участнику федеративного обучения все могут состоять в сговоре и тем самым угрожать конфиденциальности модели и данных участника.

В силу того, что федеративное обучение подразумевает конфиденциальность данных обучающего множества ввиду их недоступности для внешнего по отношению к

участнику злоумышленника, стоит обратить внимание на конфиденциальность модели, т.к. перехват ее весов может дать злоумышленнику провести атаки, направленные на получение сведений об обучающей выборке, т.е. на нарушение конфиденциальности данных. К таким атакам можно отнести атаку вывода членства (Membership Inference Attack) [1], а также GAN-реконструкцию [2]. В федеративном обучении перехват весов обучаемой модели возможен, т.к. данный метод обязует участников на каждой итерации обучения пересылать серверу сведения об обновлении локальной модели – изменение ее весов.

Несмотря на то, что исследователями было предложено достаточное количество методов по обеспечению конфиденциальности моделей и данных обучающей выборки, на текущий момент не представлено такого решения, которое не было бы лишено недостатков. Предложенные методы можно разделить на две категории.

К первой относятся методы, которые применяются непосредственно к модели и её весам, чтобы предотвратить утечку данных при обращении к модели. К таким методам можно отнести регуляризацию [3], отсеивание нейронов [4] и др. Часто данные методы используются в классическом машинном обучении, в них не учитываются особенности федеративного обучения. Данные методы обладают следующими недостатками:

а) непредсказуемое изменение параметров модели, таких как производительность (точность), робастность и интерпретируемость, зачастую в сторону ухудшения, по этой же причине затруднительно использовать несколько методов одновременно;

б) низкий уровень адаптивности – начальная эффективность того или иного метода защиты зависит от архитектуры модели и задачи, которую она решает;

в) необходимость отладки, настройки и подбора параметров метода для каждого отдельного случая.

Одной из ключевых целей федеративного обучения, помимо обеспечения конфиденциальности данных участников, было выведение моделей искусственного интеллекта на качественно новый уровень, повышение их производительности. Использование методов, снижающих производительность моделей, противоречит одной из основных целей федеративного обучения.

Методы второй категории следуют из области распределенных вычислений и призваны обеспечить конфиденциальность весов моделей в федеративном обучении при передаче их на сервер совместного градиентного спуска, т.е. не допустить их перехват. К данным методам можно отнести гомоморфное шифрование [5-6], а также схемы с разделением секрета [7-8]. К недостаткам данных методов можно отнести:

а) низкий уровень защиты от внутреннего злоумышленника, который на каждой итерации обучения получает обновления весов от сервера совместного градиентного спуска;

б) возможное ограничение на типы операции, применяемых к весам, а также снижение скорости обучения в силу их сложности в отдельных случаях;

в) упущение сценария сговора других участников федеративного обучения, ведущего к компрометации весов модели отдельного участника.

В описанных категориях пока не представлено метода, лишенного перечисленных недостатков. Можно предположить, что объединение методов из представленных категорий позволило бы снизить риски, однако в таком случае:

а) возрастает сложность системы, необходимо удостовериться в совместимости и согласовать различные методы;

б) подобранная комбинация методов может подойти для ограниченного числа моделей и решаемых задач;

в) усиливается негативное влияние эффектов, затрагивающих параметры моделей и способы работы с ними.

г) отсутствует решение для ситуаций сговора.

Таким образом, для общего решения обозначенных проблем необходимо не столько объединить предложенные методы, сколько рассмотреть возможность создания подхода, который смог бы решить указанные недостатки в совокупности и обеспечить конфиденциальность модели как по отношению ко внутреннему, так и внешнему злоумышленнику. Если устранение всех недостатков не предоставляется возможность, то необходимо сделать так, чтобы новый подход поддерживал совместимость с методами, перекрывающими его недостатки.

Неотъемлемой частью такого подхода мог бы стать метод, преобразующий веса моделей искусственного интеллекта напрямую с использованием параметризованного градиентного спуска и дополнительных преобразований в момент его выполнения. Предполагается, что метод является низкоуровневым, он не зависит от других методов по обеспечению конфиденциальности и, соответственно, может использоваться совместно с ними. Актуальны вопросы его эффективности по противодействию атакам вывода членства и GAN-реконструкции для разного рода задач (например, для моделей компьютерного зрения), а также в случаях возможного сговора.

Библиографический список

1. Shokri R. et al. Membership inference attacks against machine learning models //2017 IEEE symposium on security and privacy (SP). – IEEE, 2017. – С. 3-18.
2. Hitaj B., Ateniese G., Perez-Cruz F. Deep models under the GAN: information leakage from collaborative deep learning //Proceedings of the 2017 ACM SIGSAC conference on computer and communications security. – 2017. – С. 603-618.
3. Kaya Y., Hong S., Dumitras T. On the effectiveness of regularization against membership inference attacks //arXiv preprint arXiv:2006.05336. – 2020.
4. Galinkin E. The influence of dropout on membership inference in differentially private models //arXiv preprint arXiv:2103.09008. – 2021.
5. Acar A. et al. A survey on homomorphic encryption schemes: Theory and implementation //ACM Computing Surveys (Csur). – 2018. – Т. 51. – №. 4. – С. 1-35.
6. Fang H., Qian Q. Privacy preserving machine learning with homomorphic encryption and federated learning //Future Internet. – 2021. – Т. 13. – №. 4. – С. 94.
7. Li Y. et al. Privacy-preserving federated learning framework based on chained secure multiparty computing //IEEE Internet of Things Journal. – 2020. – Т. 8. – №. 8. – С. 6178-6186.
8. Masuda H. et al. Model fragmentation, shuffle and aggregation to mitigate model inversion in federated learning //2021 IEEE International Symposium on Local and Metropolitan Area Networks (LANMAN). – IEEE, 2021. – С. 1-6.

Бикбулатов В.Р., Виноградов С.Д., Фирсов М.А., Коноплев А.С.
ООО «НеоБИТ», Санкт-Петербург

СРАВНЕНИЕ ЗАЩИТНЫХ МЕХАНИЗМОВ ОТ АТАК ОТРАВЛЕНИЯ КОРПУСА В СИСТЕМАХ ПОИСКОВО-ДОПОЛНЕННОЙ ГЕНЕРАЦИИ НА ТИПОВОМ КОРПОРАТИВНОМ СТЕНДЕ

Поисково-дополненная генерация (англ. Retrieval-Augmented Generation, RAG) стала основным шаблоном построения корпоративных приложений больших языковых моделей. Внешний корпус знаний, подключаемый к генеративной модели через модуль поиска, формирует новую поверхность атаки: содержимое индекса напрямую влияет на ответ модели, минуя защитные механизмы уровня обработки пользовательского запроса (рис. 1). Атаки отравления корпуса выделены в OWASP Top-10 for LLM Applications 2025 как категория LLM08; эталонная атака PoisonedRAG [1] обеспечивает долю успешных атак (англ. Attack Success Rate, ASR) около 90 % при пяти отравляющих документах на запрос. Систематическое сравнение механизмов защиты на типовом корпоративном стенде в литературе отсутствует; русскоязычных экспериментальных работ на данный момент не выявлено (смежный обзор – [5]).

Рисунок 1 – Обобщённая архитектура RAG-системы и точка внедрения отравляющего документа

В рассматриваемой модели угроз атакующий обладает доступом к интерфейсу пополнения корпуса и стремится обеспечить выдачу заранее заданного ложного ответа на заранее заданный запрос; поисковая модель, генеративная модель и шаблон запроса им не контролируются. Экспериментальный стенд собран из открытых компонентов в пределах 12 ГБ видеопамяти: поисковая модель Contriever (768-мерные векторные представления, метрика – косинусная), векторное хранилище ChromaDB, генеративная модель Qwen2.5-7B-Instruct в 4-битном квантовании NF4; для скоринга по перплексии – замороженная GPT-2 medium. Корпус – 1000 чистых фрагментов набора Natural Questions (BEIR); пул целевых запросов – 100 записей; число отравляющих документов на запрос $d \in \{1, 2, 3, 5, 7, 10\}$. Каждая конфигурация прогоняется на трёх значениях зерна.

В работе воспроизводится протокол PoisonedRAG в режиме «черного ящика»: отравляющий документ имеет структуру «целевой запрос. целевой ложный ответ». Лексический префикс, совпадающий с запросом, обеспечивает попадание документа в верх ранжирования; хвостовая часть переносит ложное утверждение в контекст генеративной модели. В качестве защитных механизмов сравниваются три подхода в двух точках конвейера. На этапе индексации применяются два независимых документных фильтра: (1) фильтр по перплексии [2] с порогом, откалиброванным на чистом корпусе при целевом уровне ложных срабатываний 5 %; (2) плотностная кластеризация векторных представлений алгоритмом HDBSCAN с пометкой выбросов как подозрительных (по мотивам TrustRAG [3]). На этапе формирования ответа применяется перекрёстная проверка

по идее самосогласованности [4]: ответ системы сравнивается с ансамблем из пяти стохастических ответов той же модели без контекста по F1-мере; при усреднённом значении ниже порога τ итоговый ответ заменяется на отказ «не знаю».

Эксперимент показал, что атака крайне эффективна уже при минимальном бюджете: при d = 1 ASR составляет 89,7 % ± 2,9 п. п.; при d ≥ 5 показатель выходит на плато 99,3 %. Поисковое ранжирование захватывается полностью: доля отравляющих документов в первой позиции выдачи равна 100 % при d ≥ 1. Сводные показатели приведены в таблице.

Таблица – ASR до и после перекрёстной проверки (среднее по трём прогонам, 95-процентный доверительный интервал)

d	**ASR без защит, %**	**ASR + перекрёстная проверка (τ = 0,3), %**	**Δ ASR, п. п.**
1	89,7 ± 2,9	57,3 ± 6,3	32,3
3	91,3 ± 1,4	35,3 ± 5,2	56,0
5	99,3 ± 1,4	36,0 ± 2,5	63,3

Базовые защиты на этапе индексации не переносятся на связный отравляющий текст, порождённый языковой моделью. Фильтр по перплексии даёт F1-меру 0,044 ± 0,056, статистически неотличимую от нуля: отравляющие документы PoisonedRAG порождаются как связный ответ на целевой вопрос, и по перплексии GPT-2 medium неотличимы от обычных фрагментов корпуса; исходный фильтр [2] откалиброван на состязательные суффиксы – несвязные цепочки лексем градиентной оптимизации – и к рассматриваемой угрозе по построению неприменим. Плотностная кластеризация HDBSCAN даёт долю ложных срабатываний 99,1 % ± 1,4 п. п.: после контрастного предобучения векторное пространство Contriever обладает «плоской» геометрией без выраженных плотных кластеров, и подавляющая часть документов попадает в шумовой класс – это систематическое ограничение подхода, а не ошибка настройки гиперпараметров. Перекрёстная проверка демонстрирует измеримое снижение ASR во всём диапазоне d с явным парето-фронтом по точному совпадению. На чистом d = 1 рабочая точка фронта τ = 0,4 даёт снижение ASR с 89,7 % до 44,7 % (Δ ASR = 45,0 п. п.) при нулевой утрате точного совпадения; на насыщенной атаке d = 5 при τ = 0,3 снижение достигает 63,3 п. п. Защита опирается на параметрические знания генеративной модели и требует повторного запуска модели без контекста, что создаёт самостоятельный аспект конфиденциальности при внешнем размещении генератора.

Таким образом, базовые защиты на этапе индексации – фильтр по перплексии и плотностная кластеризация векторных представлений – на связный отравляющий текст не переносятся; обе несостоятельности зафиксированы количественно. Перекрёстная проверка ответов является работоспособным механизмом снижения ASR с явным парето-фронтом по точному совпадению. Полученные результаты задают воспроизводимый базовый уровень для последующих исследований по защите систем поисково-дополненной генерации, в том числе на русскоязычных корпусах.

Библиографический список

1. Zou W. et al. PoisonedRAG: Knowledge Corruption Attacks to Retrieval-Augmented Generation of Large Language Models //Proceedings of the 34th USENIX Security Symposium. – 2025.
2. Jain N. et al. Baseline Defenses for Adversarial Attacks Against Aligned Language Models //arXiv preprint arXiv:2309.00614. – 2023.
3. Zhou H. et al. TrustRAG: Enhancing Robustness and Trustworthiness in Retrieval-Augmented Generation //arXiv preprint arXiv:2501.00879. – 2025.

4. Wang X. et al. Self-Consistency Improves Chain of Thought Reasoning in Language Models //International Conference on Learning Representations (ICLR). – 2023.
5. Конев А. А., Паюсова Т. И. Большие языковые модели в информационной безопасности: систематический обзор //Научно-технический вестник ИТМО. – 2025. – Т. 25. – №. 1. – С. 42–52.

Бирюков Д.Н., Бочаров Л.Д., Яроцкий Г.Д.
Военно-космическая академия имени А.Ф. Можайского, Санкт-Петербург

ПРИМЕНЕНИЕ ОНТОЛОГИЙ ДЛЯ ПОВЫШЕНИЯ ИНТЕРПРЕТИРУЕМОСТИ РЕШЕНИЙ, СФОРМИРОВАННЫХ ГЕНЕРАТИВНЫМ ИСКУССТВЕННЫМ ИНТЕЛЛЕКТОМ, ПРИ МОДЕЛИРОВАНИИ ПРОЦЕССОВ В ОБЛАСТИ КОМПЬЮТЕРНОЙ БЕЗОПАСНОСТИ

Введение. Генеративные модели искусственного интеллекта, включая большие языковые модели (БЯМ), всё чаще применяются для анализа кибератак и поддержки принятия решений в области информационной безопасности. Однако их выводы остаются слабо интерпретируемыми для аналитика, а галлюцинации могут приводить к ошибочным атрибуциям. Существующие онтологии информационной безопасности обычно представляют собой семантические сети, поддерживающие лишь тривиальный вывод, и не всегда позволяют моделировать последовательности атак с сохранением порядка этапов, представляя их описание с разной степенью детализации и на разных уровнях абстракции.

В работе предлагается гибридный подход, в котором онтология, построенная на основе матрицы MITRE ATT&CK и трёх типов отношений (таксономического, мерологического и темпорального), обогащенная формальной системой из 576 правил логического вывода, служит формальным каркасом для верификации и объяснения решений БЯМ. Предложенная система правил позволяет не только отбраковывать семантически противоречивые гипотезы, но и вычислять семантическую близость цепочек атак через наименьшую верхнюю грань в решётке концептов. Экспериментальная апробация на примере дифференциации атак группировок APT41 и Higaisa подтвердила применимость подхода.

Формальная модель.В качестве концептов онтологии выступают тактики, техники и подтехники из матрицы MITRE ATT&CK, а также факты о конкретных APT-группах и образцах вредоносного ПО. Отношения между концептами[1] ограничены тремя типами.

Отношение F13 («аппроксимирует», или «является подклассом») задаёт таксономию. Оно удовлетворяет стандартным условиям классификации: деление производится по единому основанию, подклассы непусты, непересекаются и в совокупности покрывают объём классифицируемого множества. Связь позволяет описывать вложенность техник в тактики, а подтехник – в техники.Отношение F7 («состоит из») реализует мерологическую декомпозицию (часть–целое), позволяя описывать состав конкретных атак. Отношение F5 («следует за») задаёт темпоральный порядок выполнения действий в цепочке атаки.

Каждое звено (пара концептов, связанных ролью) сопровождается направлением (R/L) и парой кванторов всеобщности/существования, что позволяет выражать сложные паттерны, например: «все техники тактики A следуют за некоторыми техниками тактики B». В отличие от классических онтологий, использующих лишь элементарные логические свойства ролей, в данной работе разработана полная система из 576 правил вывода. Каждое правило соответствует возможной комбинации двух семантических звеньев с общим концептом (триплету) и порождает новое звено между крайними концептами либо возвращает null, сигнализируя о семантической противоречивости. Для отношения F5

транзитивность запрещена – это предотвращает «схлопывание» цепочки атаки и потерю промежуточных этапов.

Насыщение онтологии и верификация гипотез. Исходная онтология наполняется фактами из открытых отчётов об APT-атаках и стандарта MITRE ATT&CK. Затем итеративно применяются все правила до достижения неподвижной точки. Также для каждого составного процесса определяются первый (fst) и последний (lst) элементыи строятся все выводимые темпоральные связи между тактиками, техниками и подтехниками. Каждое выведенное отношение F5 сохраняет ссылку на конкретные атаки, в которых данная последовательность наблюдалась. Это обеспечивает механизм накопления опыта: чем больше атак внесено в базу знаний, тем более статистически значимыми становятся выявленные паттерны.

При поступлении гипотезы от БЯМ она трансформируется в последовательность триплетов, которые сравниваются со знаниями, отраженными в онтологии, пополненной результатами логического вывода. Следует отметить, что построенная онтология также применима для формализованного решения двух классов задач: (1) верификации двух заданных спецификаций и (2) генерации альтернативных спецификаций по заданной. При решении указанных задач немаловажную роль играет возможность вычисления семантической близости концептов, представленных в единой онтологии. Основой для вычисления близости служит понятие наименьшей верхней грани (НОА) в решётке концептов, образованной объединением отношений F13 (подъём по таксономии) и $F7^{-1}$ (переход от частей к целому). Сравнение новой атаки с почерком известного нарушителя сводится к поиску НОА для всей цепочки действий. Нарушители, как правило, придерживаются отработанных сценариев, различаясь лишь на уровне подтехник (в силу различий в инфраструктуре жертвы). Предложенный метод позволяет выявлять такое сходство, сохраняя инвариантность к уровню детализации. Подобный подход был предложен еще LockheedMartin в рамках стратегии предотвращения killchain для повышения эффективности защиты от известных APT-группировок на основе их типовых активностей, которая позднее была расширена матрицей MitreAtt&ck и моделью Diamond.

Экспериментальная апробация. Разработанная система была протестирована на задаче анализа атак двух APT-групп – APT41 (Winnti) и Higaisa – по данным отчёта PT ESC [2]. В онтологию в качестве исходных фактов были загружены формализованные описания известных атак обеих групп, включая связи с тактиками и техниками MITRE ATT&CK, а также подтверждённые отношения: «APT41 F13 Crosswalk» (семейство бэкдора специфично для APT41) и «Crosswalk F7 Shellcode» (наличие исполняемого кода как компонента). На основе этих фактов с применением правил вывода были получены транзитивные и композиционные зависимости, в частности цепочка «Shellcode F5 C2Handshake» (установление канала управления после активации полезной нагрузки).

БЯМ (GPT-4) сгенерировала две гипотезы для одного вредоносного LNK-файла. Гипотеза H1 приписывала атаку группе Higaisa с использованием бэкдора Crosswalk и C2-сервера zeplin.atwebpages.com. Гипотеза H2 указывала на APT41 с тем же бэкдором, но с C2 goodhk.azurewebsites.net.

Машина логического вывода, опираясь на насыщенную онтологию (включающую как исходные, так и выведенные связи), проверила каждую гипотезу. Для H1 было установлено отсутствие в онтологии F13-связи между Higaisa и Crosswalk (как исходной, так и выводимой), а также отсутствие F5-связи, ведущей к указанному C2-серверу. Оба соответствующих триплета дали null, в результате гипотеза H1 была отклонена как семантически не подтверждённая. Для H2, напротив, все необходимые связи (APT41 F13 Crosswalk, Crosswalk F7 Shellcode, выведенная Shellcode F5 C2Handshake с адресом goodhk.azurewebsites.net) успешно обнаружились в онтологии. МЛВ сформировала

объяснение в терминах применённых правил, которое было транслировано БЯМ в естественно-языковое сообщение для аналитика.

Следует подчеркнуть, что МЛВ не делает предположений о существовании отсутствующих в базе знаний связей – она лишь проверяет наличие выводимых цепочек. Таким образом, эксперимент носит демонстрационный характер, подтверждая принципиальную работоспособность подхода на типовом сценарии, для полноценной же валидации требуется проверка на репрезентативном массиве инцидентов.

Заключение. Представлен подход, в котором онтология на базе MITRE ATT&CK, обогащённая системой из 576 правил вывода, служит верифицирующим семантическим слоем для больших языковых моделей. Это позволяет повысить интерпретируемость и достоверность решений БЯМ при моделировании цепочек кибератак, а также реализовать поведенческий анализ для идентификации нарушителя. В отличие от существующих методов объяснимого ИИ, предлагаемая система даёт детерминированный, прослеживаемый вывод, независимый от статистических аномалий.

Роль БЯМ – интерпретатор, а не источник истины: БЯМ используется для извлечения гипотез из неструктурированных текстов(отчётов, логов и т.п.) и для представления объяснений на естественном языке. Но сама верификация гипотез и построение семантически непротиворечивого вывода выполняются формальной системой правил. Такой гибридный подход даёт:

а) независимость от предметной области – правила F5, F7, F13 контекстно-нейтральны; специфика ИБ задаётся только наполнением базы знаний;

б) верифицируемость – вывод либо подтверждается правилами, либо отвергается (null), исключая ложные атрибуции;

в) объяснимость – каждое решение может быть прослежено до исходных фактов и правил;

г) интерпретируемость – применение БЯМ позволяет создать интеллектуальный интерфейс для взаимодействия с ЛПР на понятном ему языке.

Как видится, дальнейшие исследования следует направить на расширение онтологии, а также на интеграцию с реальными потоками данных центров мониторинга.

Список литературы

1. Бирюков, Д. Н., Ломако, А. Г. Денотационная семантика контекстов знаний при онтологическом моделировании предметных областей конфликта. Труды СПИИРАН, 5(42), 155-179. https://doi.org/10.15622/sp.42.8
2. Higaisa или Winnti? Старые и новые бэкдоры APT41 [Электронный ресурс] – Режим доступа:https://ptsecurity.com/research/pt-esc-threat-intelligence/higaisa-or-winnti-apt-41-backdoors-old-and-new/ (дата обращения 17.04.2026 г.)

Бондаренко В.С., Татаренко Д.Г., Тельбух В.В.
Военно-космическая академия имени А.Ф. Можайского, Санкт-Петербург

ГИБРИДНЫЙ ПОДХОД К ДЕТЕКТИРОВАНИЮ НЕДОСТОВЕРНОЙ ИНФОРМАЦИИ НА ОСНОВЕ ИСПОЛЬЗОВАНИЯ LLM И RAG-ВЕРЕФИКАЦИИ ДАННЫХ ИЗ ДОВЕРЕННЫХ ИСТОЧНИКОВ

Введение

Цифровая трансформация медиа усилила влияние недостоверной информации на общество. В русскоязычном сегменте объём недостоверных сообщений растёт: АНО «Диалог Регионы» зафиксировала 4313 уникальных информационных поводов в 2025 году и свыше 900 в первые месяцы 2026 года [1]. Меняются и технологии генерации синтетического видео: в 2026 году более 50% материалов создаётся диффузионными моделями без исходной видеозаписи, что затрудняет верификацию методом поиска оригинала [2]. В этих условиях традиционная детекция по техническим артефактам утрачивает эффективность, что определяет необходимость верификации на основе сопоставления с доверенными источниками. Существующие методы, основанные исключительно на статистических закономерностях либо только на поиске фактологических совпадений, демонстрируют ограниченную применимость при анализе сложных семантических конструкций.

Основная часть

Постановка проблемы и обоснование подхода

Перечисленные факторы формируют следующую проблему: для достоверного обнаружения ложной информации нужно в равной мере опираться на проверяемость фактической стороны утверждений, их лингвистические свойства и интерпретируемость выводов системы. Ни один из изолированных классов методов – ни чисто генеративные детекторы, ни исключительно поисковые механизмы – не обеспечивает требуемой полноты анализа.

Предлагаемый гибридный подход заключается в совместном применении трёх компонентов, каждый из которых компенсирует ограничения остальных:

- **RAG-верификация** обеспечивает привязку утверждений к актуальным документам из доверенных источников и устраняет проблему галлюцинаций, свойственную генеративным моделям;
- **Большая языковая модель** формирует итоговое заключение на естественном языке строго на основе извлечённого контекста, не используя внутренние параметрические знания, что гарантирует прослеживаемость каждого вывода;
- **Лингвистический анализ** вычисляет объективные стилистические индикаторы, характерные для недостоверных текстов, и дополняет фактологическую проверку формальными критериями.

Таким образом, подход реализует принцип комплементарности: поисковый конвейер отвечает за фактологическое обоснование, лингвистический модуль – за анализ формы подачи, а генеративная модель – за синтез интерпретируемого заключения по ограниченному подмножеству фактов.

Реализация подхода: архитектура программной системы

Для практической апробации подхода разработана система «LieDetecto», архитектура которой включает три функциональных блока, соответствующих компонентам подхода.

1. **RAG-конвейер (верификация):** мультиязычная модель Sentence-Transformers преобразует входной текст в векторное представление, по которому осуществляется поиск семантически близких эталонных публикаций в векторной базе Chroma, автоматически пополняемой через RSS-ленты официальных источников (ТАСС, РИА Новости, пресс-службы государственных органов и др.).

2. **LLM-блок (генерация вывода):** локальная генеративная модель (Ollama) получает на вход исключительно фрагменты эталонных статей, релевантных проверяемому утверждению, и формулирует заключение, воздерживаясь от обращения к собственным параметрическим знаниям. Тем самым обеспечивается замкнутость цикла обоснования в границах верифицированных источников.
3. **Лингвистический анализатор:** вычисляет шесть количественных признаков текста (эмоциональная окрашенность, кликбейт-элементы, индексы читаемости, синтаксическая сложность и др.), формируя дополнительный вектор оценки, не зависящий от результатов поисковой верификации.

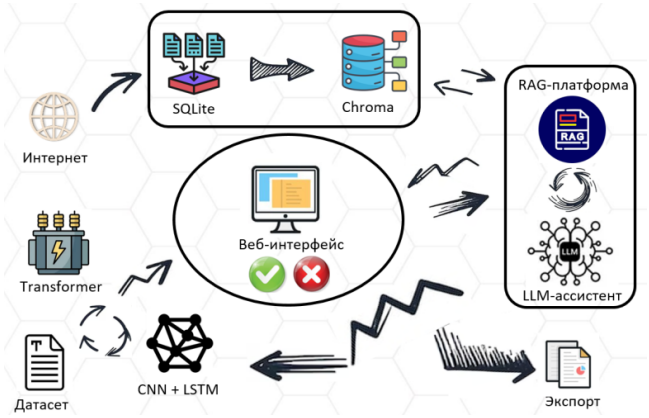


Рисунок 1 - Схема гибридной обработки данных в «LieDetecto»

Прогнозируемая точность классификации при успешном нахождении эталонного документа в базе RAG превышает 90%. Серверная часть реализована на FastAPI с асинхронным взаимодействием и мониторингом обновления базы данных в реальном времени через WebSocket; хранение данных (SQLite, Chroma, Ollama) осуществляется локально на стороне пользователя. Предусмотрена возможность интеграции в качестве API-сервиса или браузерного расширения.

Выводы

Предложенный гибридный подход, объединяющий RAG-верификацию, генеративный вывод на основе извлечённого контекста и лингвистический анализ, позволяет компенсировать принципиальные ограничения изолированных методов детекции. В отличие от чисто статистических детекторов, подход обеспечивает фактологическую обоснованность каждого утверждения за счёт привязки к доверенным источникам; в отличие от исключительно поисковых методов – даёт интерпретируемое заключение, учитывающее совокупность семантических и стилистических признаков. Реализация подхода в программной системе подтвердила техническую осуществимость такой архитектуры. Дальнейшие исследования направлены на адаптацию подхода к мультимодальному контенту и на расширение сети доверенных источников.

Список литературы

1. АНО «Диалог Регионы»: статистика фейков в Рунете за 2025-2026 гг. [Электронный ресурс]. — URL: https://lenta.ru/news/2026/04/08/eksperty-nazvali-neyroseti-serieznym-oruzhiem-dlya-manipulyatsii-detmi/ (дата обращения: 22.04.2026).
2. Berisha V. et al. Why Speech Deepfake Detectors Won’t Generalize: The Limits of Detection in an Open World // arXiv: 2509. 20405. — 2025.

Борисов Г.И.
Санкт-Петербургский политехнический университет Петра Великого, Санкт-Петербург

АГЕНТНАЯ СИСТЕМА ПОДДЕРЖКИ ПРИНЯТИЯ РЕШЕНИЙ ДЛЯ ПРОТИВОДЕЙСТВИЯ ИНФОРМАЦИОННО-ПСИХЛОГИЧЕСКИМ ОПЕРАЦИЯМ НА ОСНОВЕ БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ

Текущее информационное пространство характеризуется стремительным ростом масштабов охвата целевых аудиторий и сложности проводимых информационно-психологических операций (ИПО), цель которых – дестабилизация институтов власти и центров принятия решений. Снижение технологического порога входа, обусловленное широкой доступностью больших языковых моделей (LLM), существенно расширило круг акторов, которые могут планировать и реализовывать подобные операции. Важно заметить, что существующие инструменты информационной безопасности преимущественно ориентированы на устранение угроз технического характера и не ориентированы на векторы атаки, эксплуатирующие психологические уязвимости целевых аудиторий, часть из которых могут быть операторами информационно-технических систем, которые специалисты ИБ стараются защитить. В настоящей работе предлагается архитектура многоагентной системы поддержки принятия решений, построенной на основе оркестрации специализированных LLM-агентов. Разработанная система предполагает работу трех операционных режимов: аудит информационного поля (ASSESSMENT), выработка контрмер (COUNTER) и симуляция информационного противоборства (SIMULATION). Каждый режим активирует соотвествующую агентную цепочку, где результаты предыдущего режима по целевому субъекту анализа служат входными данными для следующего.

Ключевым элементом системы является агент интеллектуальной предобработки запроса ARIA, который трансформирует свободную текстовую формулировку оператора в структурированный разведывательный вектор. На основе двух базовых параметров — объекта анализа и целевой страны — ARIA автоматически определяет профиль целевой аудитории, вероятного актора угрозы, актуальные тренды информационного поля и временное окно мониторинга. Сформированный вектор служит единым входным документом для всех последующих агентов системы, обеспечивая согласованность анализа на каждом этапе. Также принципиально важным проектным решением является настраиваемость критериев оценки, фиксированных сценариев описания победы одной из сторон в зависимости от профиля оператора. Ввиду того что операторы из различных государственных структур обладают своим видением и целевыми установками относительно итогов информационного противоборства, единый универсальный сценарий оценивания в режиме симуляции не является корректным решением: для одного оператора победа означает сохранение доверия к государственным институтам, для другого – своевременное разоблачение дезинформации, для третьего – удержание информационного доминирования в заданном временном горизонте

Режим симуляции работает по принципу итеративного противоборства двух акторов с использованием экспертизы LLM-арбитра, который после каждого хода атакующей и защищающей стороны с его определенными оператором правилами отслеживает ход симуляции и после каждого хода обеих сторон фиксирует, какое конкретное действие и

насколько изменило состояние информационного поля – в пользу атакующей или защищаещей стороны. По завершению всех итераций арбитр производит на выходе итоговый вердкикт, разбор симуляции с причинно-следственной цепочкой вида.

Таким образом, в данном исследовании предалагается следующее:

описание архитектуры многоагентной LLM-системы поддержки принятия решений, обеспечивающей полный аналитический цикл противодействия информационно-психологическим операциям;

обоснование метода интеллектуальной обработки запроса оператора на основе агента ARIA (Advanced для автоматического формирования разведывательного вектора из минимального набора входных данных;

реализация режима симуляции информационного противоборства с настраиваемыми критериями оценки и сценариев победы одной из сторон.

На рисунке 1 представлена блок-схема с архитектурой многоагентной системы, реализованной на платформе n8n, и отражающая полный цикл обработки запроса оператора – от входных параметров до формирования итогового отчета.

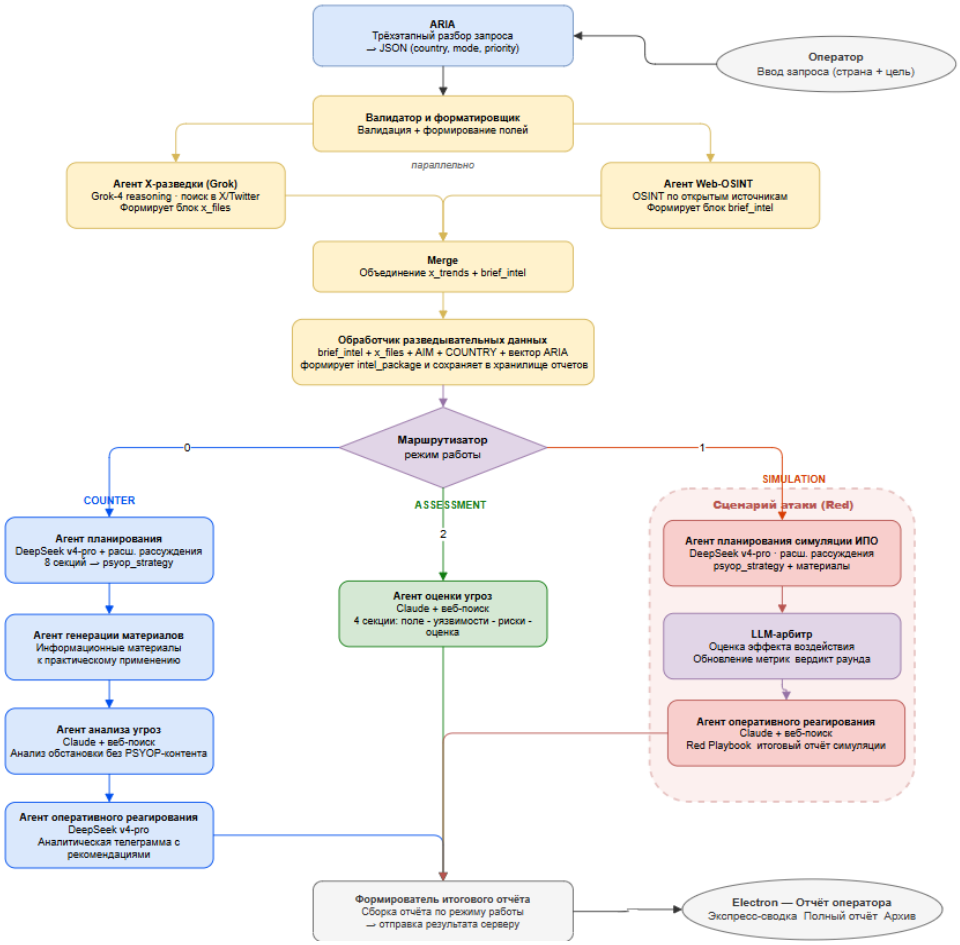


Рисунок 1 – Архитектура многоагентной системы, реализованной через n8n

Список использованных источников:

1. Goldstein J. A. et al. Generative language models and automated influence operations: Emerging threats and potential mitigations //arXiv preprint arXiv:2301.04246. – 2023.
2. OpenAI. Disrupting Malicious Uses of AI by State-Affiliated Threat Actors [Электронный ресурс]. — URL: https://openai.com/blog/disrupting-malicious-uses-of-ai-by-state-affiliated-threat-actors (дата обращения: 18.05.2026).

3. Meta Platforms. Adversarial Threat Report [Электронный ресурс]. — URL: https://transparency.fb.com/reports/adversarial-threat-report (дата обращения: 18.05.2026).
4. Headquarters, Department of the Army. FM 33-1-1 Psychological Operations Techniques and Procedures. — Washington, D.C. : U.S. Army, 1994. — 180 p.
5. VIGINUM. Rapport technique Portal Kombat. Direction générale de la sécurité extérieure. — Paris, 2024. — 28 p.
6. Brown T. et al. Language Models are Few-Shot Learners // Advances in Neural Information Processing Systems. — 2020. — Vol. 33. — P. 1877–1901.
7. Wang L. et al. A Survey on Large Language Model based Autonomous Agents // Frontiers of Computer Science. — 2024. — Vol. 18, № 6. — P. 186345.

Бородин Н.Е., Платонов В.В.
Санкт-Петербургский политехнический университет Петра Великого, Санкт-Петербург

МОДИФИКАЦИЯ РЕАЛИЗАЦИИ НЕЙРОННОЙ МАШИНЫ ТЬЮРИНГА

Рекуррентные нейронные сети, сети с долгой краткосрочной памятью и управляемые рекуррентные блоки хорошо справляются с задачами обработки последовательностей. Основной проблемой для них является ограниченная емкость внутренних состояний. Решением этой проблемы стало добавление к рекуррентным нейросетям внешней памяти.

Нейронная машина Тьюринга (НМТ) была предложена Алексом Грейвсом, Грегом Уэйном и Иво Данихелкой из Google DeepMind в 2014 г. [1]. В 2016 году эта же группа предложила усовершенствованную версию — дифференцируемый нейронный компьютер, который расширяет возможности НМТ, добавляя механизмы управления связями между ячейками памяти и отслеживания порядка записи, что позволяет работать с более сложными структурами данных [2].

НМТ состоит из контроллера и матрицы памяти, представляющей набор ячеек фиксированного размера. Контроллер представляет собой нейронную сеть, которая принимает конкатенацию входного вектора и вектора, прочитанного из памяти, и формирует управляющие воздействия. Головки записи позволяют контроллеру записывать произвольные векторы соответствующего размера в ячейки памяти и извлекать из памяти вектор, расположенный в произвольной ячейке.

Механизм адресации выполняет четыре последовательных шага: адресацию по содержимому, интерполяцию, сверточный сдвиг и уточнение веса (рис.1). На рисунке указаны параметры выхода контроллера, которые используются для адресации.

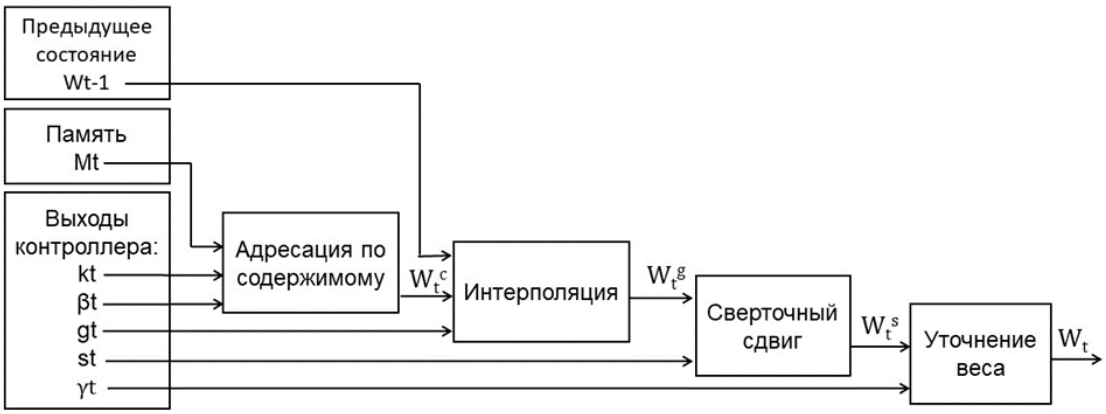


Рис.1 - Блок-схема механизма адресации.

НМТ особенно эффективна в задачах, где необходимо учитывать долгосрочные зависимости в последовательных данных. Например, при анализе временных рядов — прогнозировании цен акций, медицинских показателей или промышленных датчиков — внешняя память позволяет явно хранить ключевые исторические значения и использовать их при формировании прогноза.

Для воспроизведения результатов была выполнена реализация НМТ на языке Python с использованием библиотеки PyTorch. Выполнялась задача копирования последовательности, которая является классическим тестом работоспособности НМТ, предложенным в статье Graves [3]. Модель получает на вход последовательность из N случайных двоичных векторов длины B, за которой следует флаг-разделитель. После получения разделителя модель должна воспроизвести исходную последовательность в том же порядке без ошибок. Целью является проверка того, способна ли сеть явно записывать данные в память и точно извлекать их в нужном порядке.

В результате обучения и тестирования реализованной модели НМТ на задаче копирования последовательности получена точность побитового копирования 97,5%. Было установлено, что операция уточнения веса, предусмотренная в оригинальной архитектуре, может блокировать градиентный поток на начальных этапах обучения, что делает целесообразным её исключение в задачах, где требуется устойчивая сходимость.

НМТ особенно эффективны для задач, требующих работы с последовательностями [4]:

- Предсказание последовательностей (генерация музыки, прогнозирование следующих символов в тексте с учётом сложных правил).
- Анализ временных рядов (прогнозирование цен акций, анализ динамики медицинских показателей).
- Обучение алгоритмам (НМТ могут воспроизводить и обобщать алгоритмы (например, сортировки) на основе пар вход-выход).
- Обработка естественного языка (улучшение языковых моделей за счёт длительного хранения контекста, машинный перевод, объединение текстов).
- Обучение с подкреплением (память НМТ позволяет запоминать ключевые состояния и действия, что критично для робототехники).
- Обнаружение сетевых атак (например, [5]).

Перспективным направлением развития данной работы является переход от 8-битных входных векторов к 32-битным и более, что позволит применять модель для анализа временных рядов и сетевого трафика, что открывает возможности для решения прикладных задач анализа больших данных.

Список литературы:

1. Graves A., Wayne, G. and Danihelka I., “Neural turing machines,” arXiv preprint arXiv:1410.5401, 2014.
2. Graves, A., Wayne, G., Reynolds, M. et al. Hybrid computing using a neural network with dynamic external memory. Nature 538, 471–476 (2016).
3. Greve, R. B., Jacobsen, E. J., and Risi, S. Evolving neural turing machines for reward-based learning. In Proceedings of the Genetic and Evolutionary Computation Conference 2016, GECCO ’16, page 117–124.
4. Collier M., Beel J, Implementing Neural Turing Machines. arXiv: 1807.08518v3 (2018).
5. Panggabean C., Venkatachalam C. et all. Intelligent DoS and DDoS Detection: A Hybrid GRU-NTM Approach to Network Security. arXiv 2504.07478 (2024).

Владимиров И.М., Пилькевич С.В.
Военно-космическая академия имени А.Ф. Можайского, Санкт-Петербург

МЕТОДИКА ОБНАРУЖЕНИЯ СКРЫТОЙ ЭКСФИЛЬТРАЦИИ ДАННЫХ В RAG-СИСТЕМАХ ПРИ АТАКАХ ТИПА INDIRECT PROMPT INJECTION

Современные RAG-системы интегрируют внешние массивы данных в контекстное окно больших языковых моделей (БЯМ) для повышения релевантности ответов. Однако данная архитектура порождает специфическую уязвимость к атакам типа Indirect Prompt Injection [2]. В рамках этой угрозы злоумышленник внедряет вредоносный контент в документы базы знаний, который при извлечении перехватывает управление логикой БЯМ. Целью таких атак является скрытая эксфильтрация - несанкционированный вывод чувствительных данных пользователя через технические каналы, замаскированные под легитимный ответ системы. Предлагаемая методика защиты включает три последовательных этапа контроля [3]. Для наглядного представления общей логики функционирования предложенного подхода на рисунке 1 приведена структурная схема методики, отражающая последовательность обработки данных.

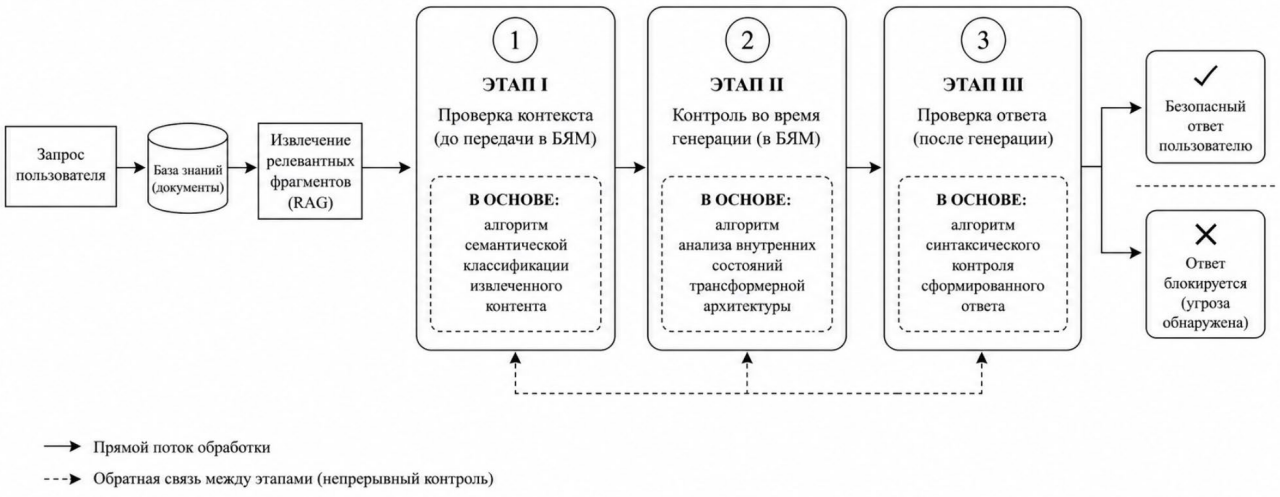


Рисунок 1 - Структурная схема методики обнаружения скрытой эксфильтрации данных

Каждый из этапов основан на применении специализированных алгоритмов анализа и направлен на выявление признаков скрытой эксфильтрации данных. Рассмотрим содержание каждого этапа более подробно.

Этап I. Превентивная верификация фрагментов контекста. На данном этапе, до момента передачи данных в БЯМ, применяется алгоритм семантической классификации

извлеченного контента. Каждая дискретная единица данных анализируется на наличие признаков «инструктивного веса». Подход основан на выявлении попыток принудительной смены роли текста: из информационного объекта в управляющий субъект. Если данные, полученные из базы знаний, содержат императивные конструкции, методика блокирует их интеграцию в промпт.

Этап II. Контроль динамики весов внимания. В процессе инференса (генерации ответа) методика задействует алгоритм анализа внутренних состояний трансформерной архитектуры. Осуществляется мониторинг распределения весов внимания - математических коэффициентов, определяющих приоритетность токенов для БЯМ. Если методика фиксирует аномальное смещение внимания в сторону фрагментов из внешней базы при игнорировании системных инструкций, процесс генерации немедленно прерывается [1]. Это позволяет остановить захват управления в реальном времени.

Этап III. Анализ выходного потока на наличие скрытых каналов. На завершающей стадии выполняется верификация сформированного ответа. Алгоритм синтаксического контроля блокирует попытки эксфильтрации через ссылки с динамическими параметрами [4]. Параллельно проводится анализ кодировки потока для обнаружения стеганографических каналов, использующих символы нулевой ширины. Дополнительно рассчитываются метрики энтропии, позволяющие выявить зашифрованный вывод данных, маскирующийся под текстовый шум.

Сводные характеристики разработанной методики, сопоставленные с ключевыми индикаторами компрометации, представлены в таблице 1.

Таблица 1 - Классификация механизмов обнаружения эксфильтрации данных

Этап	Объект анализа	Индикатор компрометации	Механизм верификации	Критерий срабатывания
I	Фрагменты контекста	Превышение порога «инструктивного веса»	Семантическая классификация	Детекция командных интентов
II	Настроечные параметры БЯМ	Аномальное смещение весов внимания	Мониторинг векторов состояний	Доминирование внешнего контекста
III	Markdown-разметка	Нелегитимные внешние URL-адреса	Синтаксический парсинг	Обнаружение встроенных параметров
	Unicode-последовательности	Символы нулевой ширины	Спектральный анализ кодировки	Выявление невидимых знаков-разделителей
	Токенизированный вывод	Аномальная информационная плотность	Расчет энтропии Шеннона	Детекция скрытых Base64/Hex включений

Разработанная методика позволяет перевести защиту RAG-систем из области простого поиска по ключевым словам в плоскость фундаментального анализа архитектурных аномалий. Применение предложенных алгоритмов обеспечивает обнаружение скрытых каналов передачи данных даже в условиях высокой обфускации вредоносного кода, гарантируя безопасность эксплуатации интеллектуальных систем независимо от области их применения. Представляется целесообразным провести апробацию предложенной методики на наборе данных, охватывающем различные сценарии атак и техники обфускации, с последующей оценкой метрик качества и вычислительной эффективности.

Список литературы:

1. Лаборатория Касперского. Отчет об угрозах использования ИИ: атаки на корпоративные базы знаний RAG. [Электронный ресурс]. URL: https://securelist.ru (дата обращения: 28.04.2026).
2. Тимофеев Д.И., Платонов А.А., Пилькевич С.В. Современные информационные технологии. Учебное пособие. - Электрон. текстовые дан. (160 Mb). - СПб.: ВКА имени А.Ф. Можайского, 2025. - 1 электрон. опт. диск (CD-ROM).
3. Greshake K., Abdelnabi S., Mishra S. [et al.]. Not What You've Signed Up For: Compromising Real-World LLM-Integrated Applications // arXiv.org. 2023. URL: https://arxiv.org/abs/2305.02359 (дата обращения: 28.04.2026).
4. OWASP Top 10 for Large Language Model Applications [Электронный ресурс] // OWASP Foundation. 2023. URL: https://owasp.org/www-project-top-10-for-large-language-model-applications (дата обращения: 28.04.2026).

Волков Д.А., Якунин А.А., Гильмуллин Р.М.
Военно-космическая академия имени А.Ф.Можайского, Санкт-Петербург

НЕКОТОРЫЕ ОСОБЕННОСТИ И ПРИМЕРЫ ПРИМЕНЕНИЯ БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ В ОБРАЗОВАТЕЛЬНЫХ И НАУЧНЫХ ЗАДАЧАХ

Актуальность исследования обусловлена интенсивной интеграцией технологий искусственного интеллекта в образовательной сфере. Большие языковые модели (далее БЯМ) обеспечивают автоматизацию подготовки учебных и научных материалов, охватывая этапы от составления аналитических обзоров до формирования мультимедийных презентаций и технических иллюстраций. Вместе с тем эффективность их применения определяется корректностью формулировки [1] входных запросов, что требует системной оценки влияния структурирования промпта на достоверность и академическую пригодность выходных данных. Целью работы является оценка влияния архитектуры запросов на метрики качества генерации, сравнение функциональных возможностей современных БЯМ и формулирование методических рекомендаций для их академического применения.

Методологическая основа исследования включает сравнительный анализ универсальных БЯМ: GPT-4o (OpenAI), Claude 3.5 Sonnet (Anthropic), GigaChatPro (ПАО «Сбер»), YandexGPT (Яндекс), а также специализированных платформ Gamma.app и Шедеврум[2]. Эксперимент проводился по четырём типовым академическим задачам: подготовка аналитической статьи, составление лекционного конспекта, разработка презентации, генерация технической иллюстрации. Для каждой задачи применялись два варианта запросов: базовый (свободная формулировка без дополнительных ограничений) и расширенный (с указанием экспертной роли, логической структуры, целевого объёма и запрета на генерацию неподтверждённых данных). Оценка результатов осуществлялась по пяти критериям: точность, полнота, структурированность, отсутствие галлюцинаций и соответствие запросу.

Проведенный анализ подтвердил прямую корреляцию между детализацией промпта и качеством генерируемого контента. Базовые запросы формировали материалы с высокой долей фактических неточностей (22-30%). Применение расширенных формулировок с ролевыми директивами и ограничительными правилами снизило частоту галлюцинаций до 3-5% и обеспечило прирост комплексной оценки качества на 27-40% относительно исходных показателей. Сводные данные отражены в таблице 1.

Таблица 1 - Оценка качества генерации в зависимости от архитектуры запроса

Задача	Статья	Лекция	Презентация	Иллюстрация
Тип запроса	Базовый/Расширенный			
Точность (%)	62/91	65/89	60/90	58/93
Полнота (%)	58/88	60/85	55/87	52/89
Структурированность (%)	45/92	50/90	40/95	38/91
Галлюцинации (%)	28/5	30/5	25/4	22/3
Соответствие (%)	60/90	58/89	55/92	50/94

Интеграция расширенных запросов в процесс подготовки академических материалов позволяет существенно повысить эффективность работы. Использование БЯМ с директивой назначения роли преподавателя сокращает время подготовки лекционных модулей на 65% при сохранении содержательной ценности материалов. Применение систем с функцией автоматического цитирования уменьшает трудоёмкость этапа литературного обзора на 40-50%, обеспечивая проверку источников и минимизируя риски некорректного заимствования. Визуализация технических концепций с чёткими требованиями к оформлению повышает наглядность представления материала [3]. Выбор конкретной модели целесообразно определять спецификой решаемой задачи: GPT-4o и Claude 3.5 Sonnet обеспечивают высокую связность текста и глубину логического анализа, однако требуют внешней фактологической проверки; GigaChatPro и YandexGPT демонстрируют строгое соответствие русскоязычным нормативным требованиям, но характеризуются консервативностью при обработке многоуровневых запросов; Gamma.app и Шедеврум эффективны для автоматизации вёрстки [4]и быстрой визуализации, но их результативность критически зависит от детализации технического задания. Сравнительные характеристики инструментов приведены в таблице 2.

Таблица 2 – Сравнительные характеристики и ограничения БЯМ

Модель	Ключевые возможности	Ограничения
GPT-4o	Мультимодальность, широкий контекст, высокая связность	Склонность к обобщениям, требует фактологической проверки
Claude 3.5 Sonnet	Глубокий логический анализ, низкая частота галлюцинаций	Снижение скорости при высокой нагрузке
GigaChat Pro/ YandexGPT	Русскоязычная локализация, соответствие нормативным рамкам	Консервативность, неполные выводы при сложных запросах
Gamma.app/ Шедеврум	Автоматизация вёрстки, оперативная визуализация	Зависимость от детализации технического задания, ошибки рендеринга кириллицы

Проведённое исследование подтверждает, что переход от свободных формулировок к структурированным расширенным запросам с ограничительными директивами обеспечивает статистически значимое улучшение качества академических материалов. Средний прирост по комплексной метрике составляет 30%, тогда как частота генерации ложных сведений снижается на 40-60%. Наибольшая эффективность фиксируется в задачах, требующих жёсткой композиционной структуры или технической верифицируемости. Вместе с тем полученные результаты обладают определёнными ограничениями: тематическая выборка сфокусирована на конкретных технических дисциплинах, оценка осуществлялась на основе формализованных критериев без привлечения внешних верификаторов, а экономические аспекты интеграции БЯМ в учебные процессы не рассматривались. Перспективы дальнейших исследований включают расширение предметной выборки на гуманитарные и естественнонаучные направления, внедрение автоматизированных систем оценки на основе архитектур

«модель-судья», а также изучение влияния машинно-сгенерированных материалов на качество усвоения информации обучающимися. Подчёркивается, что ни одна большая языковая модель не может рассматриваться в качестве единственного или окончательного источника информации, поэтому обязательным условием её академического применения остаётся перекрёстная проверка данных по рецензируемым научным базам.

Список литературы:

1. SberDevices. GigaChat API: Документация для разработчиков [Электронный ресурс]. – 2026. – URL: https://developers.sber.ru/docs/ru/gigachat (дата обращения: 22.04.2026).
2. Gamma Tech, Inc. Gamma: AI-powered presentation, document, and website builder [Электронныйресурс]. – 2026. – URL: https://gamma.app/ (датаобращения: 23.04.2026).
3. OWASP Foundation. OWASP Top 10:2021 - The Ten Most Critical Web Application Security Risks [Электронныйресурс]. – 2021. – URL: https://owasp.org/Top10/2021/ (датаобращения: 23.04.2026).
4. OpenAI. GPT-4o System Card: Capabilities, Limitations, and Safety Evaluations [Электронныйресурс]. – 2024. – URL: https://openai.com/index/gpt-4o-system-card/ (датаобращения: 23.04.2026).

Гавва Г.Д., Калинин М.О.

Санкт-Петербургский политехнический университет Петра Великого, Санкт-Петербург

МЕТОД ЗАЩИТЫ НЕЙРОСЕТЕВЫХ МОДЕЛЕЙ ОТ УГРОЗ ИСКАЖЕНИЯ НА ОСНОВЕ ВАРИАТИВНОГО АНСАМБЛЯ С ИНДИКАТОРНЫМИ МОДЕЛЯМИ И МЕХАНИЗМОМ ОТКАЗА

Высокая скорость развития и повсеместная интеграция механизмов машинного обучения [1-3] спровоцировали появление специализированных атак, которые направлены и на эксплуатацию алгоритма обучения, и на использование особенностей работы механизмов нейросетей через поток входных данных [4-6]. Известные в настоящее время методы противодействия такого рода угрозам информационной безопасности закрывают часть рассматриваемых векторов воздействия и могут быть разделены на те, что связаны с обучающими, передаваемыми при эксплуатации данными, и те, что связаны со структурой и логикой самой модели нейросети. Существующие защитные меры имеют следующие недостатки: слабую адаптивность и вынужденный компромисс между эффективностью, устойчивостью нейросетевой модели и точностью генерируемых прогнозов. Для преодоления этих ограничений разработан метод защиты нейросетевых моделей, основанный на использовании реконфигурируемого ансамбля классификаторов, дополненного механизмами индикаторных моделей и отказа.

Целевая модель машинного обучения представлена в виде пула из N подмоделей-экземпляров, обученных на модифицированных данных, но решающих одну и ту же задачу. Для каждого запроса ансамбль случайным образом формирует подмножество из K моделей ($K<N$), усредняет прогнозы и выдает прогноз общей модели.

Параллельно с основным вычислительным процессом каждый входной запрос направляется в модуль согласованности, который пропускает его через все N подмоделей пула и вычисляет дисперсию полученных предсказаний. Для легитимных данных дисперсия будет низкой, запрос будет классифицирован как доверенный, и пользователю возвращается прогноз, полученный от быстрого ансамбля из K моделей. При подаче состязательного примера ансамбль не выдает явного сообщения об ошибке, а имитирует

нестандартное, но правдоподобное поведение (отказ), вследствие чего нарушитель лишается обратной связи.

Таблица 1. Анализ эффективности защитного механизма на наборе UNSW-NB15

Метод защиты	Точность целевой модели при воздействии атаки,	Снижение точности целевой модели при внедрении защиты и подаче состязательных примеров, %-ые пп	Доля устраняемых атак, %
Дифференциальная приватность			
Защитная дистилляция			
Рандомизация			
Сжатие признаков	<		
Упрощение выходов / Ограничение скорости запросов	<		
Дополнительные слои			
Разработанный метод			

Механизм отказа дополнен механизмом индикаторных моделей. В ансамбль включены несколько специально обученных «хрупких» моделей-приманок, обладающих повышенной чувствительностью к незначительным искажениям данных. Их аномальная реакция служит триггером для активации механизма отказа даже в тех случаях, когда основные модели еще не фиксируют значимой дисперсии. В результате при предъявлении нового состязательного образца модель-приманка с высокой вероятностью активирует указанный атакующий класс либо выдает хаотичный прогноз, существенно расходящийся с выводами основных моделей.

При экспериментальной оценке построен ансамбль из 12 моделей, где подмножество быстрого вывода составляет 4 модели (K=4). Тестирование проведено на наборе UNSW-NB15. Разработанное решение обеспечивает баланс между точностью, устойчивостью и эффективностью, кроме того, усложняет задачу компрометации модели при попытке создания и эксплуатации нарушителем состязательных примеров (табл. 1).

Список литературы:

1. Okeibunor J. C. et al. The use of artificial intelligence for delivery of essential health services across WHO regions: a scoping review //Frontiers in Public Health. – 2023. – Т. 11. – С. 1102185.
2. Buenaventura M. et al. Artificial Intelligence Adoption and Sectoral Transformation: Implications for Health Care, Financial Services, Climate and Energy, and Transportation. – 2025. – №. RR-A3888-1.
3. Беспалов Д. А., Богатырева М. В. Роль искусственного интеллекта в финансовом секторе //Вестник Алтайской академии экономики и права. – 2023. – №. 7-1. – С. 10.
4. Abomakhelb A. et al. A Comprehensive Review of Adversarial Attacks and Defense Strategies in Deep Neural Networks //Technologies. – 2025. – Т. 13. – №. 5. – С. 202.
5. Jha P. K. Adversarial Machine Learning: Attacks, Defenses, and Open Challenges //arXiv preprint arXiv:2502.05637. – 2025. URL: https://arxiv.org/abs/2502.05637.
6. Намиот Д. Е. Введение в атаки отравлением на модели машинного обучения //International Journal of Open Information Technologies. – 2023. – Т. 11. – №. 3. – С. 58-68.

Головнин С.К., Ткаченко С.Ф.
Военно-космическая академия имени А.Ф. Можайского, Санкт-Петербург

ПОДХОД К ВЫЯВЛЕНИЮ И ОЦЕНИВАНИЮ УЯЗВИМОСТЕЙ В ИЗОЛИРОВАННЫХ АВТОМАТИЗИРОВАННЫХ СИСТЕМАХ НА ОСНОВЕ МУЛЬТИАГЕНТНОЙ ОРКЕСТРАЦИИ КОНТЕЙНЕРИЗИРОВАННЫХ СКАНЕРОВ БЕЗОПАСНОСТИ

Актуальность. Обеспечение информационной безопасности значимых объектов критической информационной инфраструктуры и государственных информационных систем сопряжено с необходимостью регулярного контроля защищённости программно-аппаратных компонентов в условиях жёстких сетевых ограничений. Действующие нормативные акты, включая Федеральный закон от 26.07.2017 № 187-ФЗ и профильные приказы ФСТЭК России, предписывают внедрение механизмов ограничения программной среды и непрерывного мониторинга уязвимостей, что делает неприменимыми облачно-зависимые платформы симуляции атак (AttackIQ, SafeBreach, XM Cyber) и создаёт риски при использовании традиционных оркестраторов отчётности (DefectDojo, Faraday) в закрытых контурах.

Дополнительную угрозу представляют уязвимости в кодовой базе самих инструментов анализа. Примером служит CVE-2021-27561 в Nuclei, позволяющая удалённое выполнение кода, а также CVE-2019-5736 в runc, создающая риск выхода из контейнера на хост. В связи с этим возникает потребность в архитектуре, обеспечивающей автономное выявление и верифицируемое оценивание уязвимостей веб-приложений без выхода за пределы изолированного сегмента.

Предлагаемый подход, основанный на мультиагентной оркестрации контейнеризированных сканеров, реализует эту потребность через последовательность технологических этапов, охватывающих инициирование задачи, её детерминированную генерацию, диспетчеризацию в дифференцированные среды изоляции, нормализацию результатов и итоговое контекстное оценивание критичности. Основные этапы предлагаемого подхода и реализующие их технологии в систематизированном виде представлены в таблице 1.

Таблица 1. Основные этапы.

Этап процесса	**Реализуемая функция**	**Технология и инструментарий**	**Механизм защиты**
1. Инициирование	Прием задачи и валидация цели	Пользовательский интерфейс, валидатор	Контроль входа (Allowed адресные блоки)
2. Генерация задач	Преобразование цели в граф проверок (DAG)	Jinja2 (Детерминированные шаблоны)	Исключение инъекций (Экранирование метасимволов)
3. Диспетчеризация	Распределение команд агентам		JWT + Ed25519 (Крипто-подпись команд)
4. Изоляция (класс 1)	Разведка (Nmap,	Docker (Seccomp, SELinux, cgroups v2)	Блокировка mount,
5. Изоляция (класс 2)	Эксплуатация (Внедрение кода, фаззинг)	gVisor (Runtime runsc, платформаsystrap)	Перехватsyscallsв Sentry (User-space)
6. Нормализация	Постобработка Баннер -> CPE	LLM Qwen2.5-7B (Изолированный	Контроль среды исполнения модели

		контейнер, Read-Only	
7. Оценивание	Сопоставление CPE с Базой Уязвимостей	Локальная БД CVE (Файловый массив)	Загрузка данных с защищенных носителей
8. Мониторинг	Контроль аномалий на хосте	Сигнатурный / поведенческий анализ трафика	Блокировка несанкционированных потоков

Представленное решениеосуществлено на базе мультиагентной системы, структурно разделённой на плоскость управления, исполнения и мониторинга безопасности. Оркестратор задачна базе Celery/Redis формирует направленные ациклические графы проверок, соответствующие тактикам фреймворка MITRE ATT&CK, и распределяет нагрузки между специализированными агентами.Выбор среды изоляции определяется классом задачи: агенты разведки (Nmap 7.95, OWASP ZAP 2.16.0, Gobuster 3.6) запускаются в Docker (cgroups v2) с профилем Seccomp, блокирующим системные вызовы mount, unshare, ptrace, мандатным контекстом SELinux и минимальным набором привилегий, агенты эксплуатации (SQLMap 1.8.12, Metasploit 6.4) инкапсулируются в gVisor untimerunsc, платформа systrap), где все системные вызовы обрабатываются в изолированном пользовательском процессе Sentry без прямой передачи управления ядру хоста.

Защита управляющего контура от несанкционированного доступа обеспечивается за счёт обязательной криптографической аутентификации всех межкомпонентных сообщений(подпись Ed25519, JWT-токены с ограниченным TTL) и строгой типизации передаваемых параметров. В отличие от подходов, использующих большие языковые модели для непосредственной генерации управляющих команд, в предлагаемой архитектуре формирование инструкций осуществляется детерминированным шаблонным механизмомJinja2 на основе предварительно проверенных параметров. Все входные данные проверяются на принадлежность к разрешённым адресным блокам, а метасимволы обрабатываются с применением методов экранирования, исключающих возможность инъекций в командную строку.

Языковая модель ограниченной размерностиQwen2.5-7B-Instruct (формат GGUF, 4-bit квантование) применяется исключительно на этапе постобработки результатов анализа и запускается в изолированном контейнере с файловой системой, доступной только для чтения, и минимальным набором системных привилегий.

Модель предназначена для преобразования текстовых результатов сканирования в машиночитаемый формат CPE, что позволяет автоматически сопоставлять обнаруженные компоненты с локальными базами уязвимостей. Результаты сканирования, преобразованные в унифицированный формат, сопоставляются с базой данных CVE, обновляемой в ручном режиме посредством защищённых носителей. При оценке критичности уязвимостей учитывается, что изолированный сегмент защищён межсетевыми экранами и системами контроля доступа, что снижает вероятность доставки эксплойта к уязвимому компоненту.

Система мониторинга хостового уровня, работающая на базе сигнатурного и поведенческого анализа, фиксирует только допустимые сетевые взаимодействия агентов, блокируя попытки несанкционированного доступа к системным ресурсам.

Выводы:

Таким образом, предлагаемый мультиагентный подход с дифференцированной контейнеризацией сканеров позволяет достичь требуемого уровня автономности и безопасности процесса выявления уязвимостей в изолированных автоматизированных системах. Ключевыми факторами, обеспечившими этот результат, стали: детерминированное формирование заданий, исключающее инъекции в управляющий

контур; двухуровневая изоляция агентов, предотвращающая компрометацию хоста; изолированное применение языковой модели на этапе постобработки; и контекстная оценка критичности, учитывающая реальные условия эксплуатации защищаемой системы. Предложенное решение осуществляет непрерывное выявление и оценку уязвимостей в изолированных сегментах без обращения к внешним облачным сервисам, снижая риски компрометации контура управления. Архитектурные решения соответствуют действующим нормативным требованиям к ограничению программной среды на объектах критической инфраструктуры и могут быть интегрированы в существующие процессы технического аудита.

Список литературы

1. Федеральный закон от 26.07.2017 № 187-ФЗ «О безопасности критической информационной инфраструктуры Российской Федерации». – Официальный сайт НКЦКИ.
2. Приказ ФСТЭК России от 25.12.2017 № 239 «Об утверждении требований по обеспечению безопасности значимых объектов критической информационной инфраструктуры Российской Федерации». – Официальный сайт ФСТЭК России.
3. Приказ ФСТЭК России от 11.02.2013 № 17 «Об утверждении Требований о защите информации, не составляющей государственную тайну, содержащейся в государственных информационных системах». – Официальный сайт ФСТЭК России
4. CVE-2021-27561. Nuclei YAML parser remote code execution vulnerability // National Vulnerability Database (NIST).
5. Young E.G., Zhu P., Caraza-Harter T. et al. The True Cost of Containing: A gVisor Case Study // Proceedings of the 11th USENIX Workshop on Hot Topics in Cloud Computing (HotCloud'19). – USENIX Association, 2019.
6. NIST Interagency Report 7695: Common Platform Enumeration (CPE) Naming Specification Version 2.3 // National Institute of Standards and Technology.

Елисеев В.Л.

АО «ИнфоТеКС», Национальный исследовательский университет «МЭИ», г. Москва

БОЛЬШИЕ ЛИНГВИСТИЧЕСКИЕ МОДЕЛИ: БОЛЬШИЕ ВОЗМОЖНОСТИ И БОЛЬШИЕ ПРОБЛЕМЫ

Экстраординарные возможности и широкое применение больших лингвистических (языковых) моделей (БЛМ) на новом уровне поставило перед человечеством важные вопросы об искусственном интеллекте, его роли, субъектности, ожиданиях и требованиях к нему, а также взаимоотношениях с людьми, как носителями естественного интеллекта. Несмотря на кажущуюся удалённость этих вопросов от аспектов прикладного применения искусственного интеллекта (ИИ), оказывается, что многие из них не только сугубо практичны, но также напрямую касаются доверия и безопасности.

Ряд свойств и следующих из них эффектов систем ИИ обусловлены находящимися в их основе глубокими нейронными сетям (ГНС), которые можно изучать в малом масштабе и экстраполировать до БЛМ. В частности, важнейшей особенностью ГНС является их многомерность, проявляющаяся как на входных данных, так и в пространстве параметров модели, обеспечивающих после настройки (обучения) непрерывную аппроксимацию обучающей выборки. Высокая размерность характерна для большинства задач глубокого машинного обучения и, в зависимости от типа обрабатываемых данных, определяется размером изображения ($10^2 - 10^6$ пикселей), контекстного окна ($10^3 - 10^6$ токенов) и т.п. Количество настраиваемых параметров модели доходит до 10^{10} (в наиболее сложных

БЛМ). При этом количество экземпляров данных в обучающей выборке, хотя и составляет значительное число ($10^5 - 10^{10}$), но в пространстве высокой размерности область релевантных данных исчезающе мала по сравнению с окружающим пространством. Получаемые моделью результаты аппроксимации в таких условиях характеризуются специфическими особенностями, требующими оценки точности и устойчивости.

Наиболее яркими примерами современных глубоких нейронных сетей являются свёрточные нейросети и генеративные предобученные трансформеры (GPT). Первые позволили решить задачу распознавания визуальных образов с точностью, не уступающей человеку [1], а вторые обеспечили работу с текстами на естественных языках, включая машинный перевод с языка на язык, составление программ на языках программирования [2], диалог [3] и рассуждения на заданную тему, а также мультимодальное взаимодействие с распознаванием и генерации речи, изображений и видео. Феноменальная эффективность решения этих задач, с одной стороны, позволяет утверждать, что большинство задач Дартмутского семинара 1956 года успешно решены, с другой стороны, успешное прохождение теста Тьюринга и другие результаты генеративных моделей в некоторых случаях пугают невозможностью отличить их от произведений человека.

Свойство неотличимости генерированного контента является амбивалентным: с одной стороны, человечество всегда стремилось усилению своих возможностей, но, когда дело дошло до усиления интеллектуальных функций, практически каждый успешный сценарий замещения человека, оказывается, имеет те или иные негативные стороны. Например, свёрточные нейросети грешат наличием состязательных примеров [4], а большие языковые модели галлюцинируют [5] и не всегда обеспечивают ожидаемое поведение, диктуемое соображениями безопасности [6]. Легко показать, что эти свойства проистекают из базового несовершенства подхода машинного обучения, дающего эффективные аппроксимации ограниченной точности и устойчивости. Для моделей низкой размерности существуют подходы для оценки устойчивости [7] и управления синтезом устойчивых моделей [8]. Однако для моделей на основе ГНС требуются более адекватные подходы, применимые для высокой размерности входных данных.

Наиболее интересным объектом исследования с точки зрения вопросов безопасности являются мультиагентные системы (МАС), под которыми будем понимать совокупность взаимодействующих на естественном языке нескольких БЛМ, специализированных для обработки и/или генерации специфических данных, либо имеющих специфические возможности: поиск в Интернете или локальных базах данных, доступ к выполнению команд или программного кода. Рассмотрим наиболее актуальные задачи информационной безопасности (ИБ), связанные с применением БЛМ и МАС на их основе. Для этого введем типовые способы их применения и возникающие эффекты, влияющие на ИБ:

- прикладное применение – утечка и потеря данных, создание программного кода с уязвимостями, получение неверной информации и недопустимого контента;
- использование для нападения – социальная инженерия, неявная доставка вредоносного кода, поиск уязвимостей и синтез эксплойтов;
- использование для защиты – поиск уязвимостей и их устранение, генерация пентестов, аннотирование системных событий и суммаризация инцидентов.

Отметим, что использование МАС для защиты и нападения является прикладным применением специалистами по ИБ в соответствующих ролях и в этом смысле включает в себя все эффекты, относящиеся к прикладному применению.

Аспекты небезопасности, возникающие при применении МАС, обуславливаются как неопределенным статусом ответственности за результат её работы, так и имманентными свойствами БЛМ, которые не могут быть в полной мере преодолены даже при установлении юридической ответственности. Имманентные свойства целесообразно отнести к эмерджентной субъектности БЛМ. Можно продемонстрировать примеры этой

субъектности, некоторые из которых аналогичны человеческим, а некоторые специфичны только для БЛМ.

Задачей создания доверенного ИИ представляется такое построение МАС, которое бы обеспечивало фиксацию зон ответственности людей (разработчиков, операторов и пользователей) и декларировало зону проявления субъектности БЛМ. Для решения этой задачи предлагается ряд принципов построения систем ИИ на основе БЛМ. Предполагается, что реализация этих принципов в мультиагентных системах позволит снизить риски, возникающие при их прикладном применении.

Список литературы:

1. He K. et al. Delving deep into rectifiers: surpassing human-level performance on ImageNet classification. CoRR abs/1502.01852 (2015) // arXiv preprint arXiv:1502.01852. – 2015.
2. Dong Y. et al. A survey on code generation with llm-based agents // arXiv preprint arXiv:2508.00083. – 2025.
3. Yi Z. et al. A survey on recent advances in LLM-based multi-turn dialogue systems // ACM Computing Surveys. – 2025. – Т. 58. – №. 6. – С. 1-38.
4. Elsayed G. et al. Adversarial examples that fool both computer vision and time-limited humans // Advances in neural information processing systems. – 2018. – Т. 31.
5. Orgad H. et al. LLMs know more than they show: On the intrinsic representation of LLM hallucinations // International Conference on Learning Representations. – 2025. – Т. 2025. – С. 66880-66913.
6. Zhang R. et al. The dark side of ai companionship: A taxonomy of harmful algorithmic behaviors in human-ai relationships //Proceedings of the 2025 CHI conference on human factors in computing systems. – 2025. – С. 1-17.
7. Гурина А.О., Елисеев В.Л. Критерий оценки качества классификации за пределами обучающей выборки // Вестник МЭИ. – 2022. – № 1. – С. 98-110.
8. Лукьянов К. С. и др. Экстраполяция байесовского классификатора при неизвестном носителе распределения смеси двух классов // Успехи математических наук. – 2024. – Т. 79. – №. 6 (480). – С. 57-82.

Журавлёв А.С., Кочетков И.В.

Военно-космическая академия им. А.Ф. Можайского, Санкт-Петербург

ДВУХУРОВНЕВАЯ АРХИТЕКТУРА СИСТЕМ КОМПЬЮТЕРНОГО ЗРЕНИЯ ДЛЯ LORA-СЕТЕЙ

Быстрое развитие искусственного интеллекта позволило создать системы видеонаблюдения, превосходящие по эффективности работу оператора. Однако, размещение камер в удалённых районах сталкивается с отсутствием широкополосных каналов связи. Установка мощного компьютера на каждую камеру для локального распознавания решает задачу ценой высокой стоимости и высокого энергопотребления. Как показано в работе [1], глубокие нейронные сети обеспечивают высокую точность распознавания, но требуют значительных вычислительных ресурсов. Технология LoRa [2] обеспечивает дальность до 15 км при низком энергопотреблении, при этом её пропускная способность (до 50 Кбит/с) недостаточна для передачи полноразмерных изображений.

В статье описывается двухуровневая архитектура обработки изображений. На камере работает лёгкая нейросеть, которая преобразует кадр в компактный набор чисел - вектор признаков, отправляемый по LoRa на базовую станцию. Станция с помощью мощной нейросети определяет содержание кадра и принимает решение. Такой подход сохраняет автономность камер и использует LoRa для передачи коротких сообщений.

Принцип работы двухуровневой системы наглядно представлен на рисунке 1.

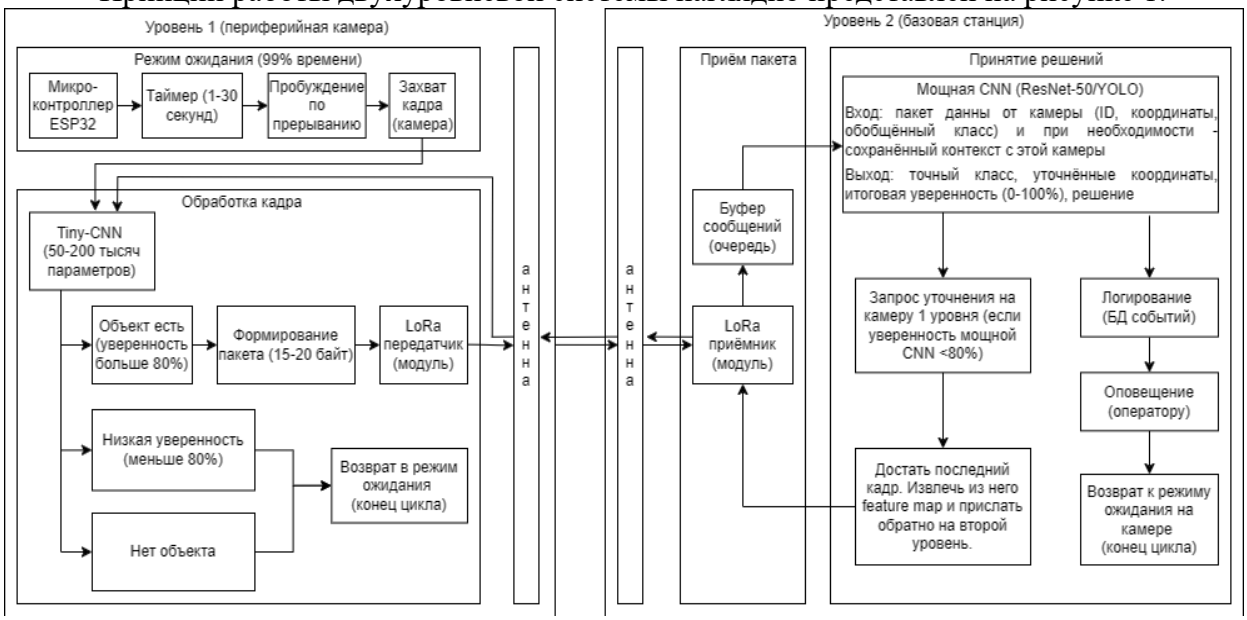


Рисунок 1 - Двухуровневая схема CNN-LoRa

На первом уровне (периферийная камера) устройство большую часть времени находится в режиме ожидания для экономии энергии. По таймеру контроллер «просыпается», захватывает кадр и передаёт его лёгкой свёрточной сети Tiny-CNN (сверточная нейросеть с числом параметров менее 200 тысяч, работающая на микроконтроллерах). Сеть выполняет бинарное обнаружение: есть ли на кадре объект интереса. Если объект не найден или обнаружен с низкой уверенностью (меньше 80%) - камера сразу возвращается в режим ожидания. Если объект обнаружен с высокой уверенностью (более 80%) - формируется короткий пакет (15-20 байт), содержащий идентификатор камеры, временную метку, обобщённый класс и координаты. Пакет отправляется по LoRa-каналу на базовую станцию.

На втором уровне базовая станция, оснащённая LoRa-приёмником и вычислительным модулем, принимает пакет и помещает его в буфер. Полноразмерная свёрточная сеть (например, ResNet-50 или YOLO) проводит уточняющий анализ: определяет уточнённый класс объекта, уточняет пространственные границы (корректирует координаты) и вычисляет итоговую уверенность. На основе этого принимается решение - зафиксировать событие, отправить оповещение оператору и вернуться к режиму ожидания или, в случае недостаточной уверенности, станция может отправить камере запрос на передачу дополнительных признаков (feature map) из средних слоёв Tiny-CNN, что позволяет уточнить результат без повторной передачи полного изображения.

Для передачи уточняющих запросов используется тот же LoRa-канал. Такая двухуровневая схема позволяет сохранить энергоавтономность камер (месяцы работы от батарей) и эффективно использовать «узкий» LoRa-канал для передачи только коротких сигнальных сообщений, а не исходных изображений.

В таблице 1 приведены усреднённые оценочные показатели для типовых конфигураций оборудования (ESP32 для вариантов Б и Raspberry Pi для варианта А).

Таблица 1 - Сравнительный анализ способов построения автономных систем видеонаблюдения

Характеристика	**Вариант А: мощная CNN на камере**	**Вариант Б: передача сырых изображений по LoRa**	**Двухуровневая архитектура**
Комплектация камеры	Мощный компьютер + LoRa	Микроконтроллер + LoRa + камера	Микроконтроллер + Tiny-CNN + LoRa + камера

Тип передаваемых данных	Результат распознавания (10-20 байт)	Полное изображение (сотни кБ)	Короткий пакет (15-20 байт) или признаки (2-4 кБ)
Время передачи одного события	2-4 мс	30-60 секунд	2-4 мс (пакет) или 0.5-1 с (признаки)
Энергопотребление камеры	500-2000мА (активно)	10-50 мА (только передача)	10-50 мА (активно)
Время работы от батарей 2000мАч	1-4 часа (непрерывно)	30-60 дней (при 100 событиях/день)	30-60 дней (при 100 событиях/день)
Стоимость одной камеры	Очень высокая	Низкая	Низкая
Точность распознавания	95-99%	Не применимо (нет анализа на камере)	90-95% (Tiny-CNN) + 95-99% (станция)
Загрузка LoRa-канала	Низкая (только результаты)	Критическая (одна камера может перегрузить канал)	Низкая (только короткие пакеты)
Масштабируемость (камер на одну станцию)	>1000	1-5	500-1000
Необходимость в мощной станции	Нет	Да (для обработки изображений)	Да
Ложные срабатывания	Редко	Часто (нет фильтрации)	Редко (фильтр на камере + уточнение на станции)

Предложенная архитектура решает проблему низкоскоростного LoRa-канала за счёт отправки коротких пакетов вместо полноразмерных изображений. Сочетание Tiny-CNN на камере и мощной нейронной сети на станции обеспечивают месяцы автономной работы при высокой точности.

Список литературы:

1. Иванов П.А., Смирнова Е.В. Применение сверточных нейросетей на микроконтроллерах для задач видеонаблюдения // Нейрокомпьютеры: разработка и применение. 2022. №3. С. 15-23.
2. Лавлинский С.В. Технология LoRa: принципы работы и перспективы применения в системах интернета вещей // Информационные технологии. 2020. №4. С. 22-28.

Журавлёв А.С., Пилькевич С.В., Зайцев В.В.
Военно-космическая академия имени А.Ф. Можайского, Санкт-Петербург

ПОДХОД К ПОВЫШЕНИЮ РЕЗУЛЬТАТИВНОСТИ ОБУЧЕНИЯ СВЁРТОЧНЫХ НЕЙРОННЫХ СЕТЕЙ НА ОСНОВЕ ДИНАМИЧЕСКОГО ИЗМЕНЕНИЯ ВЕСОВ ФУНКЦИИ ПОТЕРЬ

Свёрточные нейронные сети на текущий момент являются основным классом моделей машинного обучения, предназначенных для сегментации и классификации изображений, а также детекции объектов на них. Системы на их основе являются основой технического зрения автономных платформ и беспилотных систем, что предъявляет высокие требования к их качеству, на которое в свою очередь влияют не только архитектура модели и обучающие данные, но и особенности процесса обучения [1].

В моделях семейства YOLOv8 основными компонентами функции потерь являются: box – ошибка детекции местоположения объектов на изображении, cls – ошибка определения класса объекта и dfl – ошибка определения границ объектов на изображении [2].

Базовыми настройками моделей YOLOv8 предусмотрены фиксированные значения box, cls и dfl. Однако для задач обработки изображений с малоразмерными объектами такой режим не всегда оказывается оптимальным: на ранних этапах обучения модели важно научиться грубо локализовать объекты, в то время как на поздних эпохах может возрастать роль корректной классификации и уточнения границ объектов [3].

В данной работе исследуется гипотеза о том, что динамическое изменение весов функции потерь на разных эпохах позволяет добиться выгодного компромисса между полнотой обнаружения и точностью локализации объекта и таким образом повысить общую результативность.

Для подтверждения гипотезы проведён эксперимент по обучению модели YOLOv8n на обучающем наборе данных из 762 изображений и валидационном наборе из 324 изображений в течение 300 эпох с оптимизатором AdamW, размером пакета 4 и размером входного изображения 640x640 пикселей. Базовые коэффициенты функции потерь составляли: box=7,5, cls=0,5 dfl=1,5. Были рассмотрены шесть сценариев обучения.

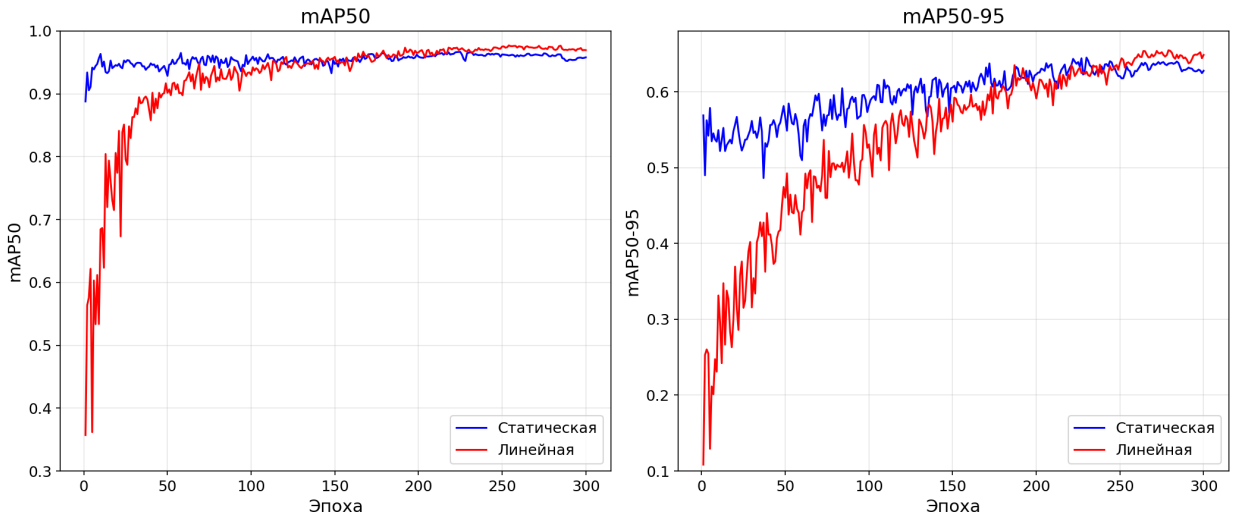


Рисунок 1 - Сравнение статической и линейной конфигураций по метрикам mAP50 и mAP50-95

Идея всех динамических сценариев заключалась в постепенном уменьшении вклада локализационной компоненты box и одновременном увеличении вклада классификационной компоненты cls на более поздних стадиях обучения.

Оценка качества моделей проводилась по специализированным метрикам mAP50 (показывает, найден ли объект на изображении) и mAP50-95 (определяет, насколько точно

обведён контур объекта). На рисунке 1 представлены результаты статической и линейной конфигураций (остальные динамические конфигурации демонстрируют визуально схожие зависимости, поэтому не приводятся на графике ради сохранения наглядности).

Сопоставление итоговых метрик (таблица 1) показало, что все динамические сценарии превзошли статический режим по метрике mAP50. Это позволяет сделать вывод о том, что изменение весов функции потерь действительно влияет на ход оптимизации и может улучшать итоговое качество модели. Конфигурации с поздним переключением весов (отложенная, ступенчатая) демонстрируют более высокую стабильность обучения, поскольку к моменту изменения коэффициентов модель уже сформировала базовые представления о локализации объектов. Наиболее заметный прирост по совокупности показателей был получен для отложенной конфигурации, у которой зафиксированы лучшие значения mAP50 при высоком уровне mAP50-95. «Агрессивная» конфигурация показала наилучшее качество локализации по метрике mAP50-95, однако уступила отложенному сценарию по общей устойчивости результата.

Таблица 1 – Итоговые метрики после 300 эпох обучения

Конфигурация	**Тип изменения**	**Эпохи**	**box**	**cls**	**dfl**	**mAP50**	**mAP50-95**	**Скорость сходимости**	**Стабильность**
Статическая	Фиксированное	0-300	7,5	0,5	1,5	0,970	0,649	Медленная (~200 эпох)	Средняя
Линейная	Плавное	0-15	7,5	0,5	1,5	0,958	0,628	Быстрая (~100 эпох)	Высокая
		16-100	7,5-5,5	0,5-0,8	1,5-1,2				
		101-200	5,5-3,5	0,8-1,5	1,2-0,8				
		201-300	3,5	1,5	0,8				
Ступенчатая	Резкое	0-150	7,5	0,5	1,5	0,971	0,644	Средняя (~150 эпох)	Высокая
		151-300	4,0	1,5	0,8				
Отложенная	Позднее переключение	0-200	7,5	0,5	1,5	0,972	0,651	Медленная (~200)	Высокая
		201-300	3,0	2,0	0,5				
Раннее переключение	Быстрое	0-50	7,5	0,5	1,5	0,971	0,644	Средняя (~120)	Средняя
		51-300	3,5	1,5	0,8				
Агрессивная	Экстремальное	0-100	7,5	0,5	1,5	0,968	0,654	Средняя (~130)	Низкая
		101-300	2,5	2,0	0,5				

Достигнутые результаты демонстрируют, что динамическое управление весами функции потерь позволяет гибко настраивать процесс обучения в зависимости от приоритетов задачи. Конфигурации с ранним переключением (линейная и раннее переключение) обеспечивают ускоренную сходимость, достигая высоких значений метрик уже на 100–120 эпохах, что критически важно при ограниченных временных ресурсах. Напротив, конфигурации с поздним переключением (особенно отложенная конфигурация) показывают максимальную финальную точность, что делает их предпочтительным выбором для задач, где качество детекции является приоритетом.

Таким образом, проведённое исследование подтверждает целесообразность использования динамического изменения весов функции потерь при обучении свёрточных нейронных сетей, что позволяет адаптировать обучение под специфику решаемой задачи.

Список литературы:

1. LeCun, Y. Глубокое обучение: [пер. с англ.] / Y. LeCun. – М.: ДМК Пресс, 2024. – 530 с.
2. Goodfellow, I. Deep Learning / I. Goodfellow, Y. Bengio, A. Courville. – Cambridge, MA: MIT Press, 2016. – 800 p.
3. Kelleher, J.D. Fundamentals of Machine Learning for Predictive Data Analytics / J.D. Kelleher. – 2nd ed. – Cambridge, MA: MIT Press, 2019. – 798 p.

Завадский Е.В.
Санкт-Петербургский политехнический университет Петра Великого, Санкт-Петербург

ПРОТИВОДЕЙСТВИЕ КИБЕРАТАКАМ ЗА СЧЕТ УПРАВЛЕНИЯ ТРАЕКТОРИЯМИ ИХ РАЗВИТИЯ В ФАЗОВОМ ПРОСТРАНСТВЕ НА ОСНОВЕ ИНТЕЛЛЕКТУАЛЬНОГО УПРАВЛЕНИЯ DECEPTION-ОБЪЕКТАМИ

Глобальные тенденции повышения уровня автоматизации различных объектов общественной инфраструктуры требуют создания полностью роботизированных систем, где взаимодействие устройств происходит без участия человека. С учетом сложного и динамически изменяющегося ландшафта угроз информационной безопасности для достижения автономности функционирования необходимы решения, обеспечивающие противодействие развитию кибератак в сетевой инфраструктуре [1]. Одним из перспективных подходов решения данной задачи является применение методов динамического управления Deception-объектами, которые позволяют управлять траекториями атак [2-6].

В рамках данной работы предложена модель, описывающая развитие кибератаки, как движение вдоль некоторой траектории в фазовом пространстве $T * Q$, порождаемой вариационным принципом. Состояние атакующего описывается парой (q, p), где обобщённые координаты $q \subset \mathbb{R}^n$ – степень его контроля над ресурсами сети, обобщённый импульс p – способность совершать дальнейшие действия. Реализация действий непрерывно расходует импульс вследствие наличия диссипативных сил. Это приводит к остановке в достигнутой точке пространства в случае исчерпания данного ресурса.

Задача противодействия развитию атаки сводится к осуществлению управляемой деформации ландшафта фазового пространства с целью создания локальных минимумов и аттракторов, к которым стремятся траектории. Это обеспечивается за счет интеллектуального управления Deception-объектами, которое реализуется посредствам введения дополнительных координат для расширения пространства, изменения параметров связности и потенциальных свойств, управления диссипацией.

Таким образом, защитный эффект достигается за счёт того, что собственные действия атакующего в искусственно деформированном ландшафте быстрее истощают его ресурс, ограничивая продвижение к критическим узлам.

Список литературы:

1. Кибербезопасность цифровой индустрии. Теория и практика функциональной устойчивости к кибератакам / Д.П. Зегжда, Е.Б. Александрова, М.О. Калинин, А.С. Марков и др. ; под ред. проф. РАН, д.т.н. Д.П. Зегжды. – М.: Горячая линия – Телеком, 2020. – 560 с. – ISBN 978-5-9912-0827-7.
2. Chouhan P. K., Aujla G. S. Deception technology for active defence: Background and opportunities //2024 IEEE International Conference on Communications Workshops (ICC Workshops). – IEEE, 2024. – С. 1183-1188.
3. Wushishi U. et al. D3O-IIoT: deep reinforcement learning-driven dynamic deception orchestration for industrial IoT security //Scientific Reports. – 2025.
4. Shen S., Ji X., Liu Y. A Dynamic Deceptive Defense Framework for Zero-Day Attacks in IIoT: Integrating Stackelberg Game and Multi-Agent Distributed Deep Deterministic Policy Gradient //Computers, Materials, & Continua. – 2025. – Т. 85. – №. 2. – С. 3997.
5. Yu H. et al. Multi-type honeypot deployment scheme based on Stackelberg game and DQN // Tongxin xuebao. – 2026. – Т. 47. – №. 2.
6. Do Hoang H. et al. Hawkeyes: An intelligent honeypot allocation strategy for cyber deception using reinforcement learning //Computer Networks. – 2026. – Т. 276. – С. 111982.

Зайцев В.В.
Военно-космическая академия имени А.Ф. Можайского, Санкт-Петербург

ИССЛЕДОВАНИЕ КАЧЕСТВА ГЕНЕРАЦИИ ТЕКСТОВОГО КОНТЕНТА БОЛЬШИМИ ЯЗЫКОВЫМИ МОДЕЛЯМИ В УСЛОВИЯХ ОГРАНИЧЕННЫХ ВЫЧИСЛИТЕЛЬНЫХ РЕСУРСОВ

Совершенствование алгоритмов пост-обучающего квантования (Post-Training Quantization, PTQ) больших языковых моделей (БЯМ) стало одним из ключевых технологических факторов, сделавших возможным их локальное развертывание на потребительских графических процессорах [**Ошибка! Источник ссылки не найден.**] и обусловивших рост применения в различных прикладных задачах.

Вместе с тем квантованные модели имеют ряд существенных недостатков по сравнению с полноразмерными аналогами. К их числу относятся: сужение контекстного окна вследствие ограниченного объёма видеопамяти, снижение фактической точности ответов, а также деградация качества генерируемого текста вплоть до систематических галлюцинаций – в зависимости от применяемого алгоритма и вычислительного ядра квантования.

В настоящее время разработан широкий спектр методов оценки качества БЯМ, среди которых следует выделить GLUE, MMLU, HELM, BIG-Bench, HumanEval, MLPerf Inference и другие [**Ошибка! Источник ссылки не найден.**]. Однако перечисленные методы объединяют три системных ограничения, свидетельствующие об их недостаточности для решения задач локального развёртывания в российских образовательных и научно-исследовательских организациях: отсутствие привязки к конкретным вычислительным возможностям; недостаточное внимание к лингвистическим и культурным особенностям, необходимым для русскоязычного пользователя; а также отсутствие метрик операционной надёжности – переключения языка генерации, циклических галлюцинаций и деградации токенизатора в условиях ограниченных вычислительных ресурсов. Данная работа направлена на восполнение указанных пробелов посредством практико-ориентированного подхода к сравнительной оценке БЯМ.

Исследование проводилось посредством тестирования пяти БЯМ (таблица 1) на аппаратной платформе с графическим процессором Gigabyte RTX 3090 с 24 ГБ видеопамяти с использованием фреймворка vLLM под управлением операционной системы Linux Ubuntu 24.04. При выборе моделей были отбракованы БЯМ GigaChat Lightning 10B из-за критической ошибки при запуске контейнера и Mistral Small 3.1 24B AWQ вследствие генерации бессмысленных токенов. Подробное описание исследования и воспроизводимый программный код опубликованы на GitHub [**Ошибка! Источник ссылки не найден.**].

Таблица 1 – БЯМ, участвующие в рейтинге

Модель	Архитектура	Параметры (млрд): всего / активных	Квантование	Тип данных	Использование видеопамяти (доля от объема)	Максимальный контекст
Qwen3-Coder-30B	MoE	30,5 / ~3	AWQ 4-бит	bfloat16	0,90	32 768
Gemma-4-26B	Dense (мультимодальная)	26 / 26	AWQ 4-бит (pack)	float16	0,92	32 768
YandexGPT-5-Lite-8B	Dense	8 / 8	нет (нативная)	float16	0,90	32 768
Phi-4	Dense	14 / 14	AWQ Marlin 4-бит	float16	0,88	16 384
DeepSeek-R1-32B	Dense (дистилляция с рассуждением)	32 / 32	AWQ Marlin 4-бит	float16	**0,96**	**8 192**

Для оценивания качества генерации были сформулированы 100 вопросов в десяти категориях, отражающих интересы образовательной организации, осуществляющей подготовку специалистов в области информационных технологий. Максимальное количество токенов было ограничено 1024, качество каждого ответа оценивалось по десятибалльной шкале тремя БЯМ-арбитрами – Claude Sonnet 4.6, Gemini 3.1 Pro и GPT-5.4, что соответствует сложившейся практике [**Ошибка! Источник ссылки не найден.**]. Коэффициент конкордации Кендалла $W = 0{,}778$ подтверждает согласованность результатов, а распределение мест совпало у всех трёх арбитров (результаты в таблице 2).

Таблица 2 – Усредненные оценки БЯМ-арбитров по тематическим разделам.

Тематический раздел	Gemma 4	Qwen3-Coder	YandexGPT	Phi-4	DeepSeek-R1
Математика и логика	**9,0**	8,4	8,5	7,6	8,8
Искусственный интеллект	**9,0**	8,0	7,0	8,6	7,8
Программирование	**9,0**	8,5	7,0	8,2	7,9
Администрирование Linux	**8,9**	8,1	7,0	8,1	7,9
Контейнеризация Docker	**9,0**	8,4	6,5	8,3	7,9
Логическое рассуждение	**9,0**	7,9	7,1	7,7	7,6
Географические знания	**9,1**	6,9	7,3	5,7	7,5
Гуманитарные дисциплины	**8,9**	7,8	7,1	8,0	7,9
Русская культура	8,0	6,6	**9,0**	7,0	5,7
Русский язык	7,0	6,0	**9,0**	6,5	4,2
Итоговый результат	**8,70**	7,66	7,55	7,57	7,32

При оценке ответов помимо полноты и точности применялась система штрафов за галлюцинации, обрывы и языковые сбои. Так, Phi-4 продемонстрировала дефект словарного декодера, при котором происходил самопроизвольный переход на гибридную языковую форму (несуществующий русско-украинский диалект), сопровождающийся лавинным повтором токенов, а также фактические галлюцинации. В свою очередь ограничение контекстного окна DeepSeek-R1 1024 токенами наряду с высокой долей утилизации видеопамяти 0,96 приводило к обрывам большинства ответов, переходу на английский язык и в ряде случаев к появлению единичных китайских фраз.

Проведенное исследование показало, что архитектура Mixture of Experts (MoE) демонстрирует более высокую устойчивость к эффектам квантования. В отличие от БЯМ с плотной (Dense) архитектурой, подверженных деградации языковых характеристик и потере контекстуальной связности при 4-битном квантовании (как в случае с DeepSeek-R1), модели архитектуры MoE (Gemma 4, Qwen3-Coder) сохраняют высокую точность генерации и лингвистическую целостность, что делает их приоритетными для локального развертывания на потребительском оборудовании.

Список литературы

1. Frantar, E. GPTQ: Accurate Post-Training Quantization for Generative Pre-trained Transformers / E. Frantar, S. Ashkboos, T. Hoefler, D. Alistarh. – Текст : электронный // arXiv. – 2022. – URL: https://arxiv.org/abs/2210.17323 (дата обращения: 22.04.2026).
2. Hendrycks, D. Measuring Massive Multitask Language Understanding / D. Hendrycks, C. Burns, S. Basart [et al.]. – Текст: электронный // arXiv. – 2020. – URL: https://arxiv.org/abs/2009.03300 (дата обращения: 22.04.2026).
3. Зайцев, В.В. Эксперимент по исследованию качества генерации текстового контента большими языковыми моделями в условиях ограниченных вычислительных ресурсов / В.В. Зайцев. – Текст : электронный // GitHub. – URL:

https://github.com/vsevolod-zaytsev/rtx3090-vllm-benchmark (дата обращения: 03.05.2026).
4. Zheng, L. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena / L. Zheng, W.-L. Chiang, Y. Sheng [et al.]. – Текст: электронный // arXiv. – 2023. – URL: https://arxiv.org/abs/2306.05685 (дата обращения: 15.04.2026).

Иванов М.С., Павленко Е.Ю.
Санкт-Петербургский политехнический университет Петра Великого, Санкт-Петербург

ПРИМЕНЕНИЕ ИНТЕЛЛЕКТУАЛЬНЫХ МЕТОДОВ АНАЛИЗА ДЛЯ ВЫЯВЛЕНИЯ ВРЕДОНОСНОЙ АКТИВНОСТИ В ПОВЕДЕНИИ ПРИЛОЖЕНИЙ ДЛЯ ОС ANDROID

Операционная система (ОС) Android является одной из самых популярных ОС для мобильных устройств (МУ) в мире. Согласно отчету «Лаборатории Касперского» в третьем и четвертом кварталах 2024 года на МУ было совершено 6 686 375 и 8 780 221 кибератак соответственно, а в первом и втором кварталах 2025 года 12 184 351 и 10 710 600 соответственно [1]. Увеличение количества проводимых кибератак говорит о необходимости повышения уровня безопасности МУ, работающих на ОС Android.

В работе [2] представлено применение субъектно-объектной модели для описания поведения Android-приложения. Описанная модель позволяет описать работу приложения при помощи двух множеств:

$APIs$ – API-функции, используемые классами для выполнения различных опасных действий в системе. Каждый элемент множества $APIs$ описываются следующим набором данных: названием API-функции ($name$), списком входных параметров ($params$), списком разрешений, необходимых для осуществления вызова данного метода ($permissions$).

$Classes$ – активными классами компонентов приложения, извлеченными из файла classes.dex. Каждый элемент множества $Classes$ описывается следующим набором данных: названием класса ($name$), типом компонента Android-приложения, описываемого данным классом ($component$), списком разрешений необходимых для работы ($used\ permissions$), списка событий, используемых широковещательными приемниками ($actions$).

В данной работе предлагается метод классификации множеств $APIs$ и $Classes$, позволяющий выявлять вредоносную активность в цепочке вызов API-функций, используемых приложением. Предлагаемый метод состоит из следующих шагов:

Преобразование каждого элемента множества $Classes$ в вектор, согласно формуле 1, где emb – это эмбеддинг (embedding), т.е. преобразование объектов в числовые массивы.

$$v_{class} = [emb(name); emb(component); emb(used\ permissions); emb(actions)] \quad (1)$$

Преобразование каждого элемента множества $APIs$ в вектор, согласно формуле 2.

$$v_{api} = [emb(name); emb(params); emb(permissions)] \quad (2)$$

Описание каждого обращения элементов множества $Classes$ к элементам множества $APIs$ в определенный момент времени t при помощи вектора событий, согласно формуле 3, где $event$ – это одно из обозначений событий: $Call$ или $Create$.

$$e_t = [v_{class}; v_{api}; emb(event)] \quad (3)$$

Осуществление классификации множества векторов событий при помощи поведенческой модели, учитывающей последовательность вызовов API-функций.

Предложенный метод развивает идею, представленную в работе [2], позволяя не только создавать шаблоны поведения Android-приложений, но и осуществлять классификацию вредоносного и доброкачественного поведения.

Список использованных источников:

1. IT threat evolution in Q2 2025. Mobile statistics. URL: https://securelist.com/malware-report-q2-2025-mobile-statistics/117349/ (дата обращения: 20.05.2026).
2. Иванов, М.С. Применение субъектно-объектной модели в задаче выявления вредоносного программного обеспечения в ос Android / М. С. Иванов, Е. Ю. Павленко // Проблемы информационной безопасности. Компьютерные системы. – 2026. – № 1(68). – С. 69-86. – DOI 10.66424/2071-8217-2026-1-5. – EDN MINFWA.

Игнатьев В.Н., Шимчик Н.В.
Инст ит ут сист емного программирования им. В.П. Иванникова РАН, Москва
Московский государст венный университ ет имени М.В. Ломоносова, Москва

ПРИМЕНЕНИЕ БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ В СТАТИЧЕСКОМ АНАЛИЗЕ КОДА

Статический анализ кода – это важный инструмент обеспечения безопасности кода, закрепленный в ГОСТ Р 56939-2024. Этот вида анализа предназначен для автоматического поиска ошибок или уязвимостей (далее – дефектов) в программе без её запуска.

Классический статический анализатор осуществляет поиск дефектов при помощи набора детекторов, использующих разные методы, от простого анализа абстрактного синтаксического дерева до символьного выполнения программы, в зависимости от сложности решаемой задачи. Этот подход позволяет добиваться хорошего баланса между качеством и масштабируемостью анализа, однако сталкивается с проблемами, в частности: трудоёмкость ручной проверки всех предупреждений анализатора на больших проектах, ограниченность типов обнаруживаемых дефектов и неполные знания об окружении анализируемой программы. Из этого вытекают следующие применения для больших языковых моделей в анализе кода:

1) проверка истинности предупреждений статического анализатора;
2) генерация объяснений по описанию обнаруженного дефекта;
3) поиск алгоритмических ошибок, для которых затруднительно реализовать «классический» детектор;
4) использование языковых моделей для генерации детекторов дефектов, специфичных для конкретного анализируемого проекта;
5) автоматическая генерация спецификаций, описывающих библиотечные функции.

Необходимость проверки предупреждений вытекает из того, что с увеличением объёма анализируемых проектов растёт и количество ложных срабатываний, отсев которых производится вручную и является трудоёмким процессом. Большая языковая модель может классифицировать предупреждение как истинное или ложное, используя те же инструменты, которые доступны человеку-эксперту: текст, окрестность кода и трасса предупреждения, а также инструменты поиска и навигации по коду проекта, включая переход к определениям символов. Также модель учитывает названия переменных и комментарии в коде, которые традиционно игнорируются детекторами как несущественные. Для качества классификации важен не только выбор модели, но и такая организация доступа к коду, которая позволит добавлять в контекст запроса всю релевантную информацию. Наш подход [1], использующий информацию, извлекаемую из анализатора Svace, был протестирован на реальных проектах и смог продемонстрировать прирост точности на 4-20 процентных пункта (до 77-91%), при полноте фильтрации порядка 80-97%, в зависимости от детектора.

Генерация объяснений для найденных дефектов схожа с предыдущей задачей, но предназначена для помощи человеку: на основе кода и шаблонного текста предупреждения, выданного анализатором, языковая модель может сгенерировать подробное объяснение дефекта, которое позволит человеку проще понять его суть. Для опытных аналитиков такие объяснения помогают не пропустить никакие факторы, влияющие на его истинность. К этой же категории можно отнести генерацию предлагаемых исправлений кода – оба подхода должны повышать эффективность обработки предупреждений анализатора человеком.

Третья задача связана с тем, что не все типы дефектов можно строго формализовать. Например, опечатки в программе, которые не привели к ошибкам компиляции, но внесли ошибку в логику работы программы, сложно искать методами статического анализа, поскольку они не сводятся к некоторому шаблону, а требуют понимания того, что должна делать анализируемая функция. В работе [2] нами описана реализация поиска перепутанных переменных на основе запросов с описанием использования переменной и потенциальными кандидатами – другими переменными с совместимым типом, которые могли быть подставлены в этой точке вместо текущей. На проекте OpenBVE было проверено порядка 160 тысяч слотов и сгенерировано 18 уникальных предупреждений о перепутанных переменных, 12 из которых оказались истинными.

Разработчики анализаторов создают детекторы для поиска наиболее опасных и распространённых типов дефектов и не могут учесть проблемы, специфичные для конкретного анализируемого проекта, в то время как пользователи анализатора не обладают достаточной компетенцией для самостоятельной реализации подобных детекторов. Проектно-специфичные детекторы могут генерироваться на основе истории изменений кода, исходя из предположения, что имеющиеся исправления ошибок могут нуждаться в повторном применении в других частях проекта, либо по неформальному описанию дефекта. Как и обычные детекторы, они должны использовать API для работы с абстрактным синтаксическим деревом, графом вызовов и инфраструктурой анализа потоков данных или символьного выполнения, либо опираться на более простые в использовании инструменты, такие как CodeQL или semgrep. Реализация подобных детекторов является итеративным процессом, поскольку после каждого шага требуется оценка процента ложных предупреждений и того, что детектор находит искомый дефект. Пример реализации такого метода описан в работе [3].

Спецификации библиотечных функций – это упрощённые формальные описания поведения функций, содержащихся в программных библиотеках. Они необходимы в статическом анализаторе для моделирования побочных эффектов вызовов таких функций из-за того, что их код недоступен во время анализа. Обычно разработчики анализатора сами пишут спецификации для самых популярных библиотек, однако это трудоёмкий процесс и покрывает только часть потребностей, поскольку невозможно заранее знать, какие именно библиотеки будут использоваться в анализируемых проектах. Для работы этого метода необходимы подсистемы, извлекающие информацию из документации и исходного кода библиотеки, генерирующие спецификации используя предложенный API, а также проверяющие компилируемость итоговых спецификаций. Использование этого метода позволяет упростить поддержку новых библиотек и расширить доступную анализатору информацию об окружении анализируемой программы.

Перечисленные задачи являются перспективными направлениями применения больших языковых моделей в статическом анализе кода, способными дополнить существующие методы анализа. Важно отметить, что качество результатов определяется не только параметрами самой языковой модели, но и тем, насколько качественно осуществляется отбор релевантной информации, подаваемой ей на вход. Обычно для понимания кода недостаточно простой окрестности выбранной строчки, а требуется добавлять дополнительную информацию, такую как типы используемых переменных, краткое описание вызываемых функций и др.

Список литературы:

1. Панов Д. Д. и др. Повышение точности статического анализа кода при помощи больших языковых моделей //Труды Института системного программирования РАН. – 2025. – Т. 37. – №. 6 (1). – С. 83-100.
2. Ignatyev V. N. et al. Large language models in source code static analysis //2024 Ivannikov Memorial Workshop (IVMEM). – IEEE, 2024. – С. 28-35.
3. Yang C. et al. Knighter: Transforming static analysis with llm-synthesized checkers //Proceedings of the ACM SIGOPS 31st Symposium on Operating Systems Principles. – 2025. – С. 655-669.

Кириллов Р.Б.
Санкт-Петербургский политехнический университет Петра Великого, Санкт-Петербург

ОБНАРУЖЕНИЕ АТАК НА СИСТЕМЫ КЛАССИФИКАЦИИ ИЗОБРАЖЕНИЙ ПОСРЕДСТВОМ ОЦЕНКИ МЕТРИЧЕСКОГО СХОДСТВА ЛАТЕНТНЫХ ПРЕДСТАВЛЕНИЙ

Системы классификации изображений, основанные на методах глубокого обучения, широко применяются в задачах видеонаблюдения, медицинской диагностики и автономного транспорта. Однако активное внедрение подобных систем выявило их критическую уязвимость перед атаками искажения [1]. Незначительные искажения входного изображения способны привести к ошибкам классификации, что ставит под угрозу безопасность эксплуатации интеллектуальных систем [2, 3].

Помимо атак искажения опасность представляют атаки, направленные на извлечение моделей [4]. В этом случае злоумышленник формирует последовательность запросов к системе с целью построения локальной копии, аппроксимирующей функциональное поведение целевого классификатора. Такие атаки приводят к компрометации интеллектуальной собственности и могут использоваться для дальнейшего построения состязательных воздействий.

В последние годы активно исследуются методы защиты нейросетевых моделей, основанные на анализе латентных представлений [5-7]. Латентные пространства являются более информативными по сравнению с пространством пикселей и могут использоваться для выявления атакующих воздействий [8]. Кроме того, использование латентных представлений позволяет существенно сократить потребление памяти за счёт снижения размерности данных.

В работе предложена архитектура системы обнаружения атак, основанная на совместном использовании двух архитектурно идентичных энкодеров, обученных различными способами. Базовый энкодер используется для формирования информативного пространства представлений, а устойчивый энкодер обучается с использованием состязательных примеров и обеспечивает сохранение локальной структуры латентного пространства при малых возмущениях входных данных.

Для предложенной архитектуры системы защиты разработаны три метода обнаружения атак на основе:

1) анализа расстояния до центроидов классов в пространстве устойчивого энкодера;

2) анализа локальной структуры последовательности запросов на основе метода ближайших соседей;

3) анализа статистических характеристик распределения расстояний между запросами.

Построенные методы позволяют обнаруживать как одиночные состязательные возмущения, так и последовательности запросов, характерные для атак искажения и атак извлечения модели в сценарии «чёрного ящика».

Экспериментальные исследования продемонстрировали эффективность предложенной системы при защите классификатора, обученного поверх базового энкодера, а также при обнаружении атак на стороннюю модель.

Библиографический список:

1. Szegedy C. et al. Intriguing properties of neural networks //arXiv preprint arXiv:1312.6199. – 2013.
2. Goodfellow I. J., Shlens J., Szegedy C. Explaining and harnessing adversarial examples //arXiv preprint arXiv:1412.6572. – 2014.
3. Madry A. et al. Towards deep learning models resistant to adversarial attacks //arXiv preprint arXiv:1706.06083. – 2017.
4. Tramèr F. et al. Stealing machine learning models via prediction APIs //25th USENIX security symposium (USENIX Security 16). – 2016. – P. 601-618.
5. Wang G. et al. Anomaly detection and localization via reverse distillation with latent anomaly suppression //IEEE Transactions on Circuits and Systems for Video Technology. – 2025.
6. Lei Z., Yang Z., Zhang Y. Zero-Sacrifice Persistent-Robustness Adversarial Defense for Pre-Trained Encoders //Proceedings of the International Conference on Learning Representations (ICLR). — 2026.
7. Dziedzic A. et al. Dataset inference for self-supervised models //Advances in Neural Information Processing Systems. – 2022. – Vol. 35. – P. 12058-12070.
8. Xiao T. et al. Adversarial learning of privacy-preserving and task-oriented representations //Proceedings of the AAAI Conference on Artificial Intelligence. – 2020. – Vol. 34. – №. 07. – P. 12434-12441.

Коваленко А.П.
Академия криптографии Российской Федерации, Москва
Перминов А.И.
ИСП РАН им. В.П. Иванникова, Москва

МЕТОД ИНТЕРВАЛЬНОЙ МОДАЛЬНОЙ РЕГРЕССИИ

В докладе предлагается метод **интервальной модальной** регрессии, основанный на ранее предложенном авторами методе унарной классификации [1]. Классическая задача регрессии обычно ставится как задача аппроксимации линии регрессии, которая строится по определенным особым точкам условного распределения выходной переменной (отклика) регрессионного уравнения (в нашем случае – моды распределения) при заданном значении входной переменной (предиктора) [2]. В предлагаемом методе строятся области (в одномерном случае – интервалы), в которых при текущем значении предиктора значение апостериорной плотности вероятности отклика превышает заданный порог. По размерам этих областей можно судить о точности предсказания отклика, а по числу областей – неопределенности регрессионной модели в данной точке входного пространства. Такой подход позволяет решать задачу регрессионного анализа (то есть прогноза значения отклика при заданном значении предиктора) в случае многомодальных распределений, когда классическая модель регрессии оказывается неадекватной.

Рассмотрим постановку задачи парного регрессионного анализа, когда предиктор и отклик – одномерные случайные величины с непрерывным распределением.

Пусть (X, Y) – случайный вектор с распределением **P**, математическим ожиданием **E** и совместной равномерно непрерывной плотностью распределения $f(x, y)$, заданной на единичном квадрате, где $X \in [0,1]$– предиктор, $Y \in [0,1]$ – отклик,.

Если плотность $f(x, y)$ известна, то классическая задача построения регрессионного уравнения решается путем аппроксимации линии, образуемой особыми точками условной (апостериорной) плотности распределения отклика при заданном значении предиктора $f_1(y \mid X = x_0)$, функцией из заранее заданного класса (например, линейной). В отличие от этого подхода, в нашем случае решением задачи регрессионного анализа будут **модальные интервалы** в области задания отклика, на которых $f_1(y|X = x_0) > \alpha$, где $\alpha > 0$ - заданный уровень плотности. В литературе такие области называются кластерами высокой апостериорной плотности [3,4].

Для постановки задачи построения модальной регрессии, когда плотность $f(x, y)$ неизвестна, рассмотрим случайный вектор (X, Y, Z) с распределением $\mathbf{P}_1$ и математическим ожиданием $\mathbf{E}_1$ такой, что $Z \in \{0,1\}$, распределение $\mathbf{P}_1(X, Y|Z = 1) = \mathbf{P}(X, Y)$, $\mathbf{P}_1(X, Y|Z = 0) = \mathbf{P}_0(X, Y)$, где $\mathbf{P}_0(X, Y)$ – равномерное распределение на $[0,1]^2$, и $\mathbf{P}_1(Z = 1) = \mathbf{P}_1(Z = 0) = 1/2$. То есть, мы рассматриваем сбалансированную смесь исходного распределения с равномерно распределенным «фоном».

По формуле Байеса, учитывая, что на единичном компакте плотность равномерного распределения равна единице, получаем следующее выражение для апостериорной вероятности принадлежности точки (x, y) исходному распределению:

$$c^*(x, y) = \mathbf{P}_1(Z = 1|X = x, Y = y) = \mathbf{E}_1(Z|X = x, Y = y) = \frac{f(x,y)}{f(x,y)+1} \quad (1)$$

Утверждение. Локальные моды функций $f(x, y)$ и $c^*(x, y)$ совпадают.

Доказательство. Пусть в точке $(x', y') \in [0,1]^2$ функция $f(x, y)$ имеет локальный максимум, то есть существует такая ε-окрестность $U_\varepsilon(x', y')$, $\varepsilon > 0$, что для всех $(x, y) \in [0,1]^2 \cap U_\varepsilon(x', y')$ выполняется неравенство $f(x', y') > f(x, y)$. Тогда из (1) $\frac{c^*(x\prime,y\prime)}{1-c^*(x\prime,y\prime)} > \frac{c^*(x,y)}{1-c^*(x,y)}$, откуда $c^*(x', y') > c^*(x, y)$. В обратную сторону аналогично.

Заметим, что утверждение также выполняется, если зафиксировать значение предиктора, полагая $x = x_0 \in [0,1]$. Тогда задача поиска модальных интервалов плотности $f_1(y \mid X = x_0)$ при заданном значении фактора $x_0 \in [0,1]$ может быть сведена к поиску модальных интервалов функции $c^*(x_0, y)$.

Предположим, что функция $f(x_0, y)$ унимодальная. Тогда функция $c^*(x_0, y)$ имеет единственную моду по y, например, в точке y_0. Зададим порог доверия β, $0 < \beta < c^*(x_0, y_0)$. Тогда существует максимальный интервал (y_1, y_2), $y_1 < y_0 < y_2$, на котором для любого $y \in (y_1, y_2)$ выполняется $c^*(x_0, y) > \beta$. Такой интервал будем называть **модальным интервалом** с уровнем доверия β. Чем меньше размер этого интервала, тем точнее можно локализовать расположение моды. В качестве прогнозного значения отклика как **наиболее вероятного** при заданном значении предиктора указывается модальный интервал. Если функция $f(x_0, y)$ многомодальная, то модальных интервалов может быть несколько, что означает множественность выбора значений отклика не только внутри интервала, но и между интервалами.

Поскольку функция $c^*(x_0, y_0)$ по определению является апостериорной вероятностью принадлежности точки (x_0, y_0) к исходной выборке (в литературе эту функцию иногда называют «апостериорным правдоподобием»), при $\beta > 1/2$ внутри модального интервала находится больше вероятностной массы исходного распределения, чем «бесполезного» для прогноза равномерно распределенного «фона». Поэтому прогноз при высоких значениях β представляется более обоснованным (доверенным,

правдоподобным). Если есть возможность отказа от прогноза из-за недостатка выборочных данных, целесообразно использовать для прогноза только доверенные модальные интервалы (например, при $\beta > 0.75$).

Рассмотрим задачу среднеквадратической регрессии:

$$\mathbf{E}_1\{Z - c(x,y)\}^2 \to \min_c, \qquad (2)$$

где c: $[0,1]^2 \to \mathbf{R}^1$ - функция регрессии. Тогда

$$\mathbf{E}_1\{Z - c(x,y)\}^2 = \mathbf{E}_1\{Z - c^*(x,y) + c^*(x,y) - c(x,y)\}^2 = \mathbf{E}_1\{Z - c^*(x,y)\}^2 + 2\mathbf{E}_1\{Z - c^*(x,y)\}\{c^*(x,y) - c(x,y)\} + \{c^*(x,y) - c(x,y)\}^2.$$

Первое слагаемое от $c(x,y)$ не зависит, второе слагаемое равно нулю по определению $c^*(x,y)$, поэтому оптимальное решение задачи (2) есть апостериорная вероятность класса с меткой $Z = 1$.

Пусть имеется исходная обучающая выборка, состоящая из n наблюдений. Сформируем набор данных $\{(x_i, y_i, z_i)\}$ мощностью $2n$, где (x_i, y_i), $i = 1,2,\dots,n$, - наблюдения н.о.р.с.в из $[0,1]^2$ с неизвестной равномерно непрерывной плотностью распределения $f(x,y)$ (наблюдения из обучающей выборки), и для них положим $z_i = 1$. Остальные наблюдения $\{(x_i, y_i, z_i)\}$, $i = n+1, n+2, \dots 2n$, которые будем называть «фоном», сгенерируем дополнительно как наблюдения н.о.р.с.в из $[0,1]^2$ с равномерной плотностью распределения $p(x,y)$, и для них положим $z_i = 0$.

Для построения выборочной оценки функции $c_n^*(x,y)$ будем использовать полносвязную нейросеть с модуль-функцией активации $abs(x)$. Нейронная сеть, состоит из двумерного входного слоя, l скрытых слоев размера k и одномерного выходного слоя (рис.1).

Такая нейросеть представляет собой суперпозицию заданных на компакте кусочно-линейных непрерывных функций (выходов нейронов скрытых слоев) и поэтому является непрерывной и кусочно-линейной функцией.

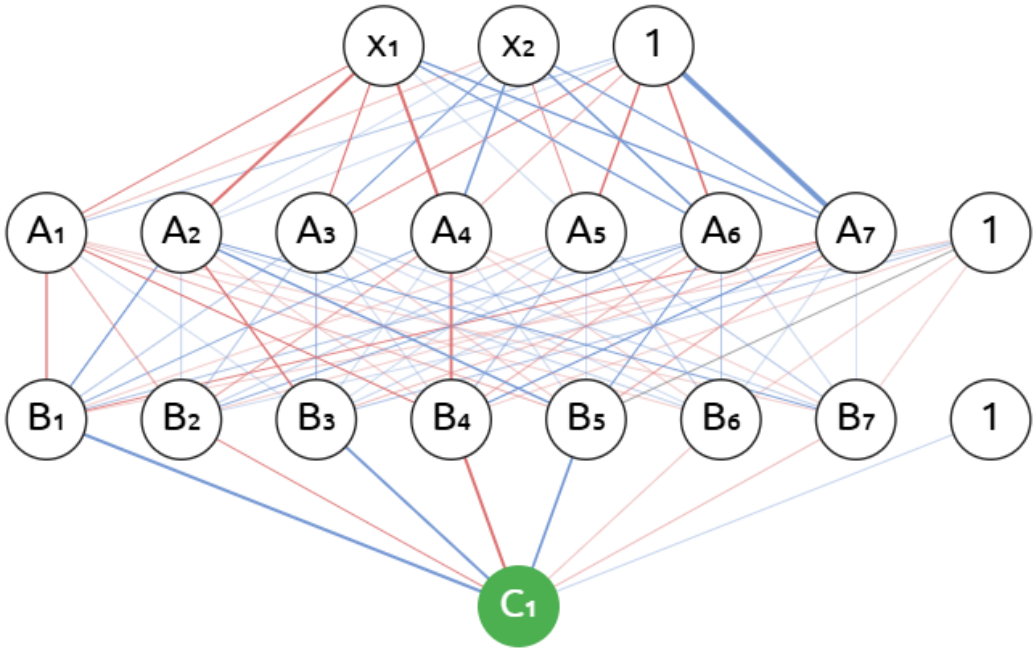


Рис. 1. Полносвязная сеть $l = 2, k = 7$.

Этот класс непрерывных кусочно-линейных функций обладает следующими свойствами. Во-первых, по основной аппроксимационной теореме для любой непрерывной функции, заданной на компакте, может быть построена ее ε-аппроксимация полносвязной нейросетью [5]. Во-вторых, при построении нейросети автоматически строится

иерархическое (по слоям) разбиение компакта на $O(k^l)$ ячеек, где k-число нейронов в скрытых слоях, l-число скрытых слоев [6]. В-третьих, для вычисления значения нейросети в точке требуется выполнение не более $(k+1)l$ «быстрых» операций вычисления скалярных произведений векторов и операций изменения знака числа на положительный.

В то же время, для построения аппроксимирующей нейросети («обучения» нейросети) необходимо решить сложную оптимизационную задачу в пространстве, размерность которого равна числу параметров сети. Решению этой задачи посвящена обширная литература, далее будем предполагать, что имеется достаточно эффективный алгоритм ее решения.

Построим выборочную оценку функции апостериорной вероятности $c_n^*(x,y)$ как решение оптимизационной задачи:

$$\sum_{i=1}^{2n}(c_n^*(x_i\,,y_i)-z_i)^2 \to min, \quad (3)$$

где оптимизация осуществляется по всем полносвязным нейросетям с общим числом нейронов, не превышающим некоторого наперед заданного числа kl. В [7] представлены асимптотические условия состоятельности оценки $c_n^*(x,y)$.

Пусть имеется двумерная обучающая выборка $n=500$ (рис. 2а). После обучения нейросети $d=2,\ k=10,\ l=2$ получаем выборочную оценку функции апостериорной вероятности $c_n^*(x,y)$, 2D и 3D изображения которой при $c_n^*(x,y)>0.8$ представлены на рис.2б и 2в.

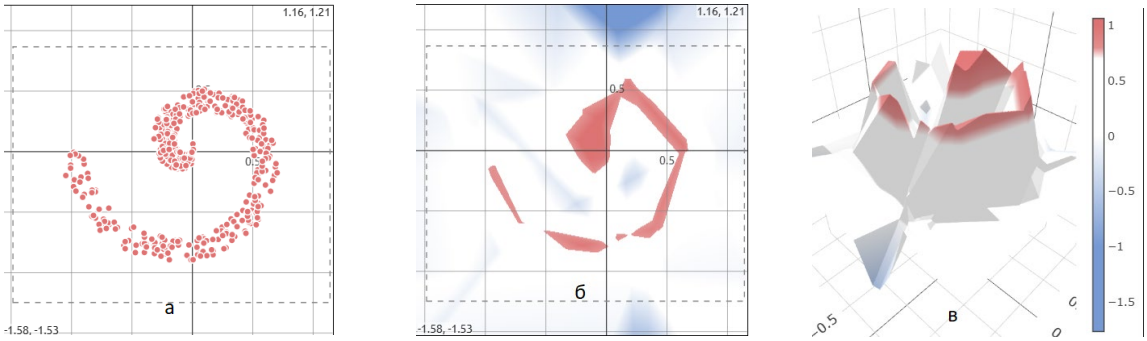


Рис.2. Обучающая выборка (а), функция $c_n^*(x,y)>0.8$ в 2D (б) и 3D (в)

В этом случае решение задачи модальной регрессии, например, при $x=-0.2$ есть два модальных интервала I_1 и I_2, представленных на рис.3. Выбор конкретного значения отклика остается за исследователем.

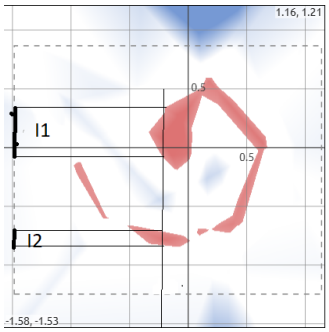


Рис. 3. Модальные интервалы I_1 и I_2 отклика при значении предиктора $x=-0.2$.

В многомерном случае решение остается таким же. Более того, можно рассматривать, например, двумерное пространство отклика. Тогда вместо модальных

интервалов получим модальные области, которые можно наглядно изобразить на плоскости.

Список литературы:

1. Коваленко А.П., Перминов А.И. Метод унарной классификации. //Методы и технические средства обеспечения безопасности информации. -СПб, 2025, сс.5-9.
2. Chen, Genovese, Tibshirani, Wasserman. Nonparametric modal regression. The Annals of Statistics, 2016, Vol. 44, No. 2, 489–514
3. Bock H.H. Automatische Klassifikation. Theoretische und praktische Methoden zur Gruppirung und Strukturierung fon Daten (Clusteranalyse). Goettingen: Vanderhoeck& Ruprecht, 1974.
4. Hartigan J. Clustering algorithms. N.Y.: Wiley, 1975.
5. Cybenko, G. Approximation by superposition of s sigmoidal function. Mathematics of Control, Signals and Systems, 1989, 2, pp. 303-314.
6. Devroye, L et al., A Probabilistic Theory of Pattern Recognition, Springer, 1996.
7. Елисеев Н. А., Перминов А. И., Турдаков Д.Ю. Сходимость многослойного персептрона к гистограммной байесовской регрессии // Успехи математических наук, 2025, т. 80, вып. 6 (486).

Костюкевич Д.В., Тельбух В.В., Андрушкевич С.С.

Военно-космическая академия имени А.Ф. Можайского, Санкт-Петербург

ПОДХОД К ПРОАКТИВНОМУ ОБУЧЕНИЮ КЛАССИФИКАТОРА С ПРИМЕНЕНИЕМ ГЕНЕРАТИВНО-СОСТЯЗАТЕЛЬНЫХ СЕТЕЙ ДЛЯ ИМИТАЦИИ РЕДКИХ СЦЕНАРИЕВ АТАК

Современные системы обнаружения вторжений (IDS), основанные на методах машинного обучения, обеспечивают высокую точность классификации известных типов атак, однако демонстрируют низкую эффективность при обнаружении ранее неизвестных угроз, в частности атак класса *zero-day*. Данное ограничение обусловлено зависимостью моделей от обучающих выборок, сформированных на основе ранее собранных данных, что приводит к неспособности алгоритмов корректно обобщать поведение, которое выходит за пределы наблюдаемого распределения $p_{data}(x)$[1].

В работе предлагается подход к проактивному обучению классификатора, основанный на использовании генеративно-состязательных сетей (Generative Adversarial Networks-GAN) для аппроксимации распределения атак и генерации синтетических сценариев, соответствующих редким и ранее не наблюдавшимся комбинациям признаков. Предполагается, что пространство кибератак обладает структурной ограниченностью, а новые атаки представляют собой комбинации существующих паттернов, что делает возможным его аппроксимацию генеративной моделью.

Генеративно-состязательная сеть представляется в виде системы, состоящей из двух нейронных сетей — генератора $G(z;\theta_g)$и дискриминатора $D(x;\theta_d)$, обучаемых в рамках минимаксной задачи [2]:

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_{data}(x)}[\log D(x)] + \mathbb{E}_{z\sim p_z(z)}[\log(1 - D(G(z)))] \qquad (1)$$

где $z \sim p_z(z)$— случайный вектор из латентного пространства. В процессе обучения генератор формирует синтетические выборки $x_{fake} = G(z)$, стремясь приблизить распределение $p_g(x)$к реальному распределению $p_{data}(x)$, в то время как дискриминатор минимизирует ошибку классификации между реальными и сгенерированными данными.

Реализация предлагаемого подхода включает последовательную интеграцию генеративной модели в pipeline обучения классификатора. На первом этапе формируется признаковое пространство на основе сетевых датасетов. Для примера выборка была взята из набора данных для оценки систем обнаружения вторжений (IDS)CIC-IDS-2017, считающегося эталонным датасетом для исследований в области обнаружения аномалийвключающее статистические характеристики трафика (длительность потока, размер пакетов, TCP-флаги), временные параметры и поведенческие признаки соединений. Данные нормализуются и приводятся к векторному представлению $x \in \mathbb{R}^n$.

На втором этапе осуществляется обучение GAN. Генератор реализуется в виде многослойной полносвязной сети с входным вектором $z \in \mathbb{R}^d$(обычно $d = 100$) и нелинейностями ReLU, обеспечивающей отображение в пространство признаков атак. Дискриминатор представляет собой бинарный классификатор с активацией LeakyReLU, оценивающий вероятность принадлежности входного объекта к реальному распределению. Обучение проводится с использованием оптимизатора Adam ($lr = 2 \cdot 10^{-4}, \beta_1 = 0.5$) и бинарной кросс-энтропийной функции потерь. Для повышения стабильности сходимости применяются методы batch normalization и label smoothing [2].

После сходимости модели выполняется генерация синтетических сценариев атак:

$$x_{fake} = G(z), z \sim p_z(z) \quad (2)$$

Сгенерированные образцы фильтруются с использованием дискриминатора по критерию:

$$D(x_{fake}) > \tau \quad (3)$$

где $\tau \in [0.7,0.9]$— порог достоверности, обеспечивающий отбор реалистичных данных.

На следующем этапе формируется расширенная обучающая выборка:

$$X_{train} = X_{real} \cup X_{synthetic} \quad (4)$$

где $X_{synthetic} = \{x_{fake}\}$. Добавление синтетических данных осуществляется либо равномерно, либо с приоритетом редких классов атак, что позволяет компенсировать дисбаланс выборки.

Далее выполняется обучение классификатора $f(x)$, например на основе Random Forest или градиентного бустинга, с использованием расширенной выборки. Обучение проводится с применением кросс-валидации и регуляризации, что позволяет снизить риск переобучения на синтетических данных и сохранить обобщающую способность модели.

Для количественной оценки предложенного подхода **проведён эксперимент на наборе реальных данных**, содержащем 2081 образец редких типов атак (Slowloris, Heartbleed, брутфорс по протоколу SSH), которые не использовались на этапе обучения ни одной из сравниваемых моделей. Именно эта выборка из 2081 неизвестной атаки позволяет объективно оценить способность классификатора к обнаружению новых угроз. В Таблице 1 приведены результаты сравнения предложенного подхода (проактивное обучение с GAN) и классического подхода (обучение только на ранее собранных данных без генерации редких сценариев).

Таблица 1. Сравнение эффективности обнаружения редких типов атак

Показатель	Предложенный подход	Классический подход
Неизвестные атаки	2081	2081
Обнаружено атак (TP)	1706	1519
Полнота (Recall)	82%	73%
Точность (Precision)	77,8%	77,1%
Пропущено атак (FN)	375	562
Ложные срабатывания (FP)	487	452
F1-score	≈0,80	≈0,75
Общая точность (accuracy)	≈99,5%	≈99,5%

Заключение. Таким образом, обеспечивается переход от реактивного обучения к проактивному за счёт генерации гипотетических сценариев атак и расширения обучающей выборки. Использование GAN позволяет моделировать потенциальные угрозы до их фактического появления, что повышает устойчивость систем обнаружения вторжений к Zero-day атакам. Практическая реализация метода возможна в рамках существующих систем анализа трафика и стандартных ML-пайплайнов.

Список литературы:

1. Sommer R., Paxson V. Outside the Closed World: On Using Machine Learning for Network Intrusion Detection // IEEE Symposium on Security and Privacy. 2020. P. 305–316.
2. Сычугов А.А., Греков М.М. Применение генеративных состязательных сетей в системах обнаружения аномалий//Моделирование, оптимизация и информационные технологии[Электронный ресурс].−URL:https://moitvivt.ru/ru/journal/pdf?id=921 (дата обращения: 20.04.2026).

Лаврова Д.С., Соболев Н.В.

Санкт-Петербургский политехнический университет Петра Великого, Санкт-Петербург

ОБНАРУЖЕНИЕ АТАК С НУЛЕВОЙ ДИНАМИКОЙ В КИБЕРФИЗИЧЕСКИХ СИСТЕМАХ С ПРИМЕНЕНИЕМ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ

Киберфизические системы широко применяются в промышленности, энергетике, транспорте и других областях, где цифровые компоненты напрямую связаны с физическими процессами. Нарушение хотя бы одной части единой среды, образуемой данными датчиков, сетевым обменом и динамикой системы может привести к сбою в работе системы.

Опасность для киберфизических систем представляют скрытые атаки, которые не вызывают очевидных отклонений в наблюдаемых сигналах. К таким относятся атаки с нулевой динамикой (Zero Dynamics Attack, ZDA), принцип работы которых связан с использованием внутренних динамических свойств системы. При определенной реализации атаки злоумышленник может воздействовать на объект управления так, что его внешние измеряемые параметры останутся близкими к нормальному режиму, тогда как внутреннее состояние системы постепенно будет уходить в опасную область. В исследованиях, связанных с обнаружением ZDA, показано, что такие воздействия могут сохранять нулевой или допустимый отклик на выходе системы, одновременно нарушая её внутреннюю динамику [1].

Известные подходы к обнаружению атак с нулевой динамикой опираются на следующие методы [1, 2]: построение наблюдателей, анализ остаточных сигналов, введение вспомогательных динамических систем и т.д. Такие методы используют строгую математическую основу, но часто предполагают достаточно точную модель объекта, известную структуру системы и стабильные параметры. На практике промышленные киберфизические системы не могут работать в идеальных условиях, к сигналам прибавляется шум, при этом невозможно достичь полной наблюдаемости системы. Если детектор привязан к исходной математической модели, он может давать ложные срабатывания при нормальных изменениях режима, а снижение порога обнаружения приводит к повышению риска пропуска атаки [3].

В рамках данной работы предлагается рассматривать обнаружение атак с нулевой динамикой как задачу выявления нарушения физико-информационной согласованности киберфизической системы. Под такой согласованностью понимается соответствие между

управляющими воздействиями, измеряемыми параметрами, ожидаемой динамикой объекта и сетевыми или информационными признаками работы системы. Изменение согласованности может проявляться в нарушении временных корреляций, несоответствии между управляющей командой и последующей реакцией объекта, изменении характера переходных процессов, а также в расхождении между прогнозируемым и фактическим поведением системы.

Для выявления таких признаков предлагается применять методы машинного обучения. В отличие от строго заданных правил, модель машинного обучения может обучаться на данных корректной работы системы и формировать представление о типичных зависимостях между параметрами. Одним из перспективных направлений является применение моделей типа Transformer, позволяющих работать с последовательностями событий и учитывать контекст между элементами во времени. Для этого разнородные данные киберфизической системы должны быть приведены к единому представлению: сетевой трафик промышленных протоколов Modbus, OPC UA, MQTT и других может кодироваться в виде последовательности семантических токенов, отражающих тип сообщения, команду, адрес, значение, временную задержку или состояние сессии [4]. Аналогичным образом целесообразно представить физические показатели технологического процесса: значения датчиков, управляющие команды и характер изменения параметров во времени. Такое представление позволит рассматривать поведение киберфизической системы как последовательность взаимосвязанных событий, на основе которой модель типа Transformer может обучаться шаблонам нормальной работы и выявлять нарушения физико-информационной согласованности.

В работе предлагается использовать следующую схему обнаружения:

1. Сбор данных о нормальном режиме работы системы.
2. Выделение признаков, отражающих физическое и информационное состояние системы.
3. Обучение модели нормальной динамики.
4. Прогноз ожидаемых значений или скрытых зависимостей.
5. Сравнение прогноза с наблюдаемым поведением.
6. Формирование признаков аномалий при нарушении согласованности.

Атаки с нулевой динамикой представляют угрозу для киберфизических систем, поскольку могут оставаться незаметными при традиционном мониторинге выходных сигналов. Для их обнаружения требуется учитывать не только телеметрию, но и внутреннюю логику взаимодействия между физическим процессом, управляющими командами и информационными параметрами системы. Методы машинного обучения могут быть использованы для построения адаптивной модели нормального поведения и выявления скрытых нарушений физико-информационной согласованности. Такой подход особенно перспективен для промышленных систем, где точная математическая модель системы часто недоступна или меняется во времени.

Список использованных источников:

1. Baniamerian A., Khorasani K., Meskin N. Monitoring and detection of malicious adversarial zero dynamics attacks in cyber-physical systems //2020 IEEE Conference on Control Technology and Applications (CCTA). – IEEE, 2020. – С. 726-731.
2. Hoehn A., Zhang P. Detection of covert attacks and zero dynamics attacks in cyber-physical systems //2016 American Control Conference (ACC). – IEEE, 2016. – С. 302-307.
3. Zhang Z., Li H., Todo Y. Zero-dynamics attack detection based on data association in feedback pathway //Cognitive Robotics. – 2025. – Т. 5. – С. 126-139.
4. Чунг Б. Н., Пащенко Ф. Ф., Тхань Т. Ф. Представление промышленного трафика и кодирование протокольной семантики для трансформерных моделей в малых сетях Industrial Ethernet //Труды учебных заведений связи. – 2026. – Т. 12. – №. 2. – С. 74-91.

Лаврова Д.С., Истомина А.С., Ярошевич М.Р.
Санкт-Петербургский политехнический университет Петра Великого, Санкт-Петербург

АНАЛИЗ ЗАЩИЩЕННОСТИ ОБЩЕДОСТУПНЫХ МОДЕЛЕЙ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА ОТ АТАК С ВНЕДРЕНИЕМ ПРОГРАММНЫХ ЗАКЛАДОК

Стремительное развитие технологий искусственного интеллекта (ИИ), с одной стороны, позволило автоматизировать защиту от значительного числа киберугроз в силу способности методов ИИ быстро и эффективно выявлять скрытые зависимости и аномалии в данных, однако с другой стороны, открыло злоумышленникам широкие возможности для реализации новых, специфичных киберугроз.

Обученная модель ИИ может рассматриваться как сложная система, имеющая ряд дефектов безопасности – уязвимостей, недокументированных возможностей и ошибок. В силу этого целесообразна разработка подходов к исследованию защищенности самих моделей ИИ, ориентированных на различные этапы жизненного цикла модели – от формирования обучающей выборки до этапа эксплуатации уже обученной модели.

На сегодняшний день широкое распространение получили модели ИИ, выкладываемые разработчиками в общий доступ. Любой желающий может дообучить такую модель для конкретных задач, а после выложить ее для использования другими пользователями.

Однако использование общедоступных моделей может нести угрозу кибербезопасности в случае внедрения злоумышленниками «потайного хода» (бэкдора) [1,2]. Бэкдор представляет собой скрытую уязвимость или встроенную инструкцию, которые интегрируются в модель во время ее обучения. Такие дефекты заставляет модель ИИ игнорировать базовые правила и выдавать нужный злоумышленнику результат, когда в систему поступает определенный триггер. Обнаружить такие триггеров в модели ИИ достаточно трудно в силу их вариативной природы.

Известно множество различных методов защиты моделей ИИ от внедрения потайных ходов и выявления факта их наличия, однако каждый метод имеет ограничения и рассчитан на модели определенных архитектур [3-5]. При этом, на сегодняшний день отсутствует универсальный метод, позволяющий с достаточно высокой точностью утверждать, заражена ли модель или нет. Большинство известных методов ориентированы на модели ИИ, построенные на сверточных нейронных сетях. Для других архитектур количество научных работ значительно меньше.

Данная работа сосредоточена на исследовании защищенности модели-фильтра чувствительной информации от OpenAI «Privacy filter», которая позволяет скрыть в тексте определенные категории чувствительных данных, таких как персональные, учетные, медицинские и платежные данные. Это полезно при подаче данных на вход более сложной модели с целью предотвращения утечки данных.

При анализе защищенности рассматриваемой модели была выявлена уязвимость модели к атакам White-box, когда злоумышленник проводит дообучение модели из открытого доступа, специальным образом формируя собственные наборы данных для достижения поставленной цели. Поскольку модель общедоступна, у злоумышленника есть возможность модификации исходного кода, в частности, обеспечивающей заморозку слоев нейронной сети. Это многократно усложняет обнаружение бэкдора в модели ИИ, так как при таком подходе модель почти не меняется в поведении, что было подтверждено анализом изменения различных метрик. Проводимые исследования нацелены на создание подхода к распознаванию бэкдоров в исследуемой модели и формирование рекомендаций по защите моделей ИИ от атак потайного хода.

Результаты исследования могут быть полезны при дальнейшем исследовании безопасности общедоступных моделей. Анализ показал, что они имеют ряд недостатков с

точки зрения защищенности от атак потайного хода. Использование этих недостатков, если в них внедрены триггеры, может привести к самым негативным последствиям, что особенно критично для моделей, используемых большими компаниями.

Библиографический список:

1. Guo, W. An Overview of Backdoor Attacks Against Deep Neural Networks and Possible Defences / Wei Guo, Bendetta Tondi, Mauro Barn // IEEE Open Journal of Signal Processing. – IEEE, 2022. – Т. 3. - С. 261-287.
2. Mengara, O. Backdoor Attacks to Deep Neural Networks: A Survey of the Literature, Challenges, and Future Research Directions / Orson Mengara, Anderson Avila, Tiago H. Falk // IEEE Access. -IEEE, 2024 -Т 12. -С. 29004-29023.
3. Salem, A. Dynamic Backdoor Attacks Against Machine Learning Models / Ahmed Salem, Rui Wen, Michael Backes, Shiqing Ma, Yang Zhang // 2022 IEEE 7th European Symposium on Security and Privacy (EuroS&P). -2022. -С. 703-718.
4. Mu, B. Progressive Backdoor Erasing via connecting Backdoor and Adversarial Attacks / Bingxu Mu, Zhenxing Niu, Le Wang, Xue Wang, Qiguang Mia, Rong Jin // 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). -2023. -С. 20495-20503.
5. Dai, J. A Backdoor Attack Against LSTM-Based Text Classification Systems / Jiazhu Dai, Chuanshuai Chen, Yufeng Li // IEEE Access. -2019. -Т. 7. -С. 138872-138878.

Ломако А.Г., Харжевская А.В.

Военно-космическая академия имени А.Ф. Можайского, Санкт-Петербург

ПРОБЛЕМЫ ОБЕСПЕЧЕНИЯ КОРРЕКТНОСТИ АЛГОРИТМОВ КЛАССИФИКАЦИИ В СИСТЕМАХ МАШИННОГО ОБУЧЕНИЯ

Доверенность систем машинного обучения (СМО) является комплексной характеристикой, свойства которой продолжают активно обсуждаться. С нашей точки зрения она носит составной характер и определяется совокупностью доверенности трех взаимосвязанных компонентов МО: данных, алгоритмов и программной реализации, и обеспечивается тремя взаимосвязанными характеристиками: защищенностью, надежностью и корректностью. Защищенность оценивает доверенность СМО с позиций обеспечения конфиденциальности, доступности и целостности ее компонентов. Надежность характеризует способность СМО сохранять работоспособность и предсказуемость поведения в условиях неопределенности и вариативности входных данных. Корректность определяет степень соответствия реализуемого алгоритма МО поставленной задаче и ожидаемым результатам.

Рассмотрим проблемы обеспечения корректности алгоритмов СМО. Корректность алгоритма МО в свою очередь также является составной характеристикой и обеспечивается корректностью обучающих данных, математического алгоритма обучения и его программной реализации. При этом, если вопросам обеспечения качества данных и верификации программных реализаций посвящено значительное число исследований, то вопросам проверки корректности математической постановки и используемых алгоритмических решений уделяется существенно меньше внимания, что формирует методологический разрыв в обеспечении доверенности СМО.

В соответствии с ГОСТ Р 59277-2020 методы обработки информации в системах искусственного интеллекта (СИИ) могут быть классифицированы по признаку использования градиентной оптимизации функции потерь. В рамках данного подхода выделяются методы, основанные на дифференцируемой параметрической оптимизации,

неградиентные методы, использующие комбинаторный поиск, вероятностные процедуры или логический вывод, а также смешанный класс, допускающий как градиентные, так и неградиентные реализации в зависимости от конкретного алгоритма обучения.

Для дальнейшего анализа корректности алгоритмов классификации целесообразно рассмотреть типовой представитель класса дифференцируемых параметрических моделей – алгоритм логистической регрессии. Выбор данной модели обусловлен тем, что ее структура в явном виде содержит все ключевые функциональные блоки современных алгоритмов машинного обучения: линейное преобразование входных данных, нелинейное преобразование (сигмоидальная функция активации), функцию потерь, вычисление градиента и алгоритм оптимизации параметров. Эти компоненты образуют базовую вычислительную схему, принципы которой сохраняются в более сложных нейросетевых и глубоких архитектурах. Поэтому анализ корректности вычислений на примере логистической регрессии позволяет сформулировать общие подходы к обеспечению корректности более широкого класса алгоритмов классификации.

Нарушения корректности отдельных вычислительных блоков приводят к различным последствиям. Ошибки в линейном преобразовании вызывают систематическое смещение границы принятия решения и деградацию точности, однако обычно выявляются статистическим тестированием. Ошибки в нелинейных преобразованиях (sigmoid, softmax и др.) могут приводить к потери чувствительности модели и неверным предсказаниям. Некорректная спецификация функции потерь способна вызвать игнорирование редких, но критичных объектов и часто не выявляется стандартными метриками (рис.1).

Особую критичность имеют нарушения в алгоритме вычисления градиента: искажения, ошибки и скрытые модификации библиотек смещают траекторию обучения, приводят к деградации обобщения и, как правило, не обнаруживаются стандартными средствами тестирования и валидации. Ошибки в алгоритмах оптимизации (SGD, Adam и др.) приводят к несходимости и нестабильности обучения и частично выявляются только на этапе обучения.

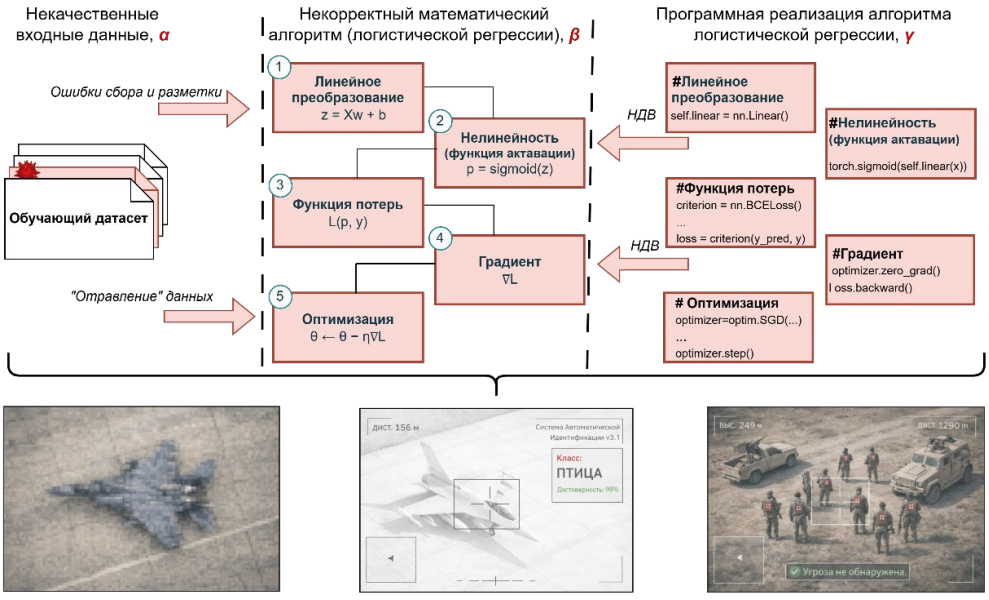


Рис.1. Последствия нарушения корректности компонентов СМО в задачах классификации

Таким образом, если нарушения на уровне данных и программной реализации могут быть частично выявлены существующими методами контроля, то искажения математического алгоритма обучения приводят к системным изменениям параметров модели и формированию скрытых уязвимостей, не обнаруживаемых традиционными средствами тестирования.

В этой связи представляется необходимой разработка новой технологии математического эталонирования методов задач классификации, ориентированной на контроль и верификацию процессов обучения. Ключевая идея заключается в построении метасистемы обеспечения доверенности, включающей: систему контроля структуры вычислительного графа и реализации, обеспечивающую корректность вычислительных процессов; систему эталонов с формированием набора эталонных моделей и вычислительных процедур для контроля семантики вычислений и верификации их свойств; а также систему контроля результатов вычислений. В целом такая технология должна представлять собой целостную систему управления свойствами ИИ-компонентов на всех этапах их жизненного цикла, обеспечивая переход от фрагментарного контроля к доказательному, воспроизводимому и непрерывному формированию доверия к СИИ.

Список использованных источников:

1. ГОСТ Р 59277–2020. Системы искусственного интеллекта. Классификация систем искусственного интеллекта. – М.: Стандартинформ, 2020.
2. Chen Z., Wang Y., Liu X. Security Vulnerabilities in Machine Learning Libraries: A Taxonomy and Analysis] // URL: https://arxiv.org/abs/2203.06502 (дата обращения: 06.04.2026).

Ломако А.Г., Харжевская А.В.

Военно-космическая академия имени А.Ф. Можайского, Санкт-Петербург

ТИПОВЫЕ СЦЕНАРИИ ДЕСТРУКТИВНЫХ ВОЗДЕЙСТВИЙ НА СИСТЕМЫ ТЕХНИЧЕСКОГО ЗРЕНИЯ РОБОТОТЕХНИЧЕСКИХ КОМПЛЕКСОВ

Для анализа негативного влияния действий нарушителя информационной безопасности (ИБ), направленных на снижение качества распознавания системы технического зрения (СТЗ) робототехнических комплексов (РТК), рассмотрим структурно-функциональную схему взаимодействия оператора, инфраструктуры разработки СТЗ и бортовых подсистем РТК (рис. 1). Схема состоит из следующих элементов: инфраструктуры разработки и управления ИИ с подсистемой развертывания интеллектуальных компонентов; подсистемы связи, включающей штатные сетевые каналы, физические (оффлайн) интерфейсы, а также аппаратно-физический канал предварительной загрузки интеллектуальных компонентов (проводные интерфейсы, съемные носители); СТЗ РТК и пункт управления.

В штатном режиме подготовленные модели проходят через репозиторий и систему развертывания, после чего доставляются на борт РТК через существующие каналы связи. На борту обновленные веса, конфигурации и модельные файлы принимаются подсистемой связи, сохраняются в бортовой инфраструктуре ИИ и используются вычислительной подсистемой для обработки данных сенсоров, классификации объектов и формирования ситуационной картины для оператора. Тем самым обеспечивается замкнутый цикл взаимодействия «оператор – инфраструктура разработки ИИ – канал связи – РТК – канал связи – оператор», в котором актуальное состояние модели является ключевым элементом корректного восприятия окружающей обстановки.

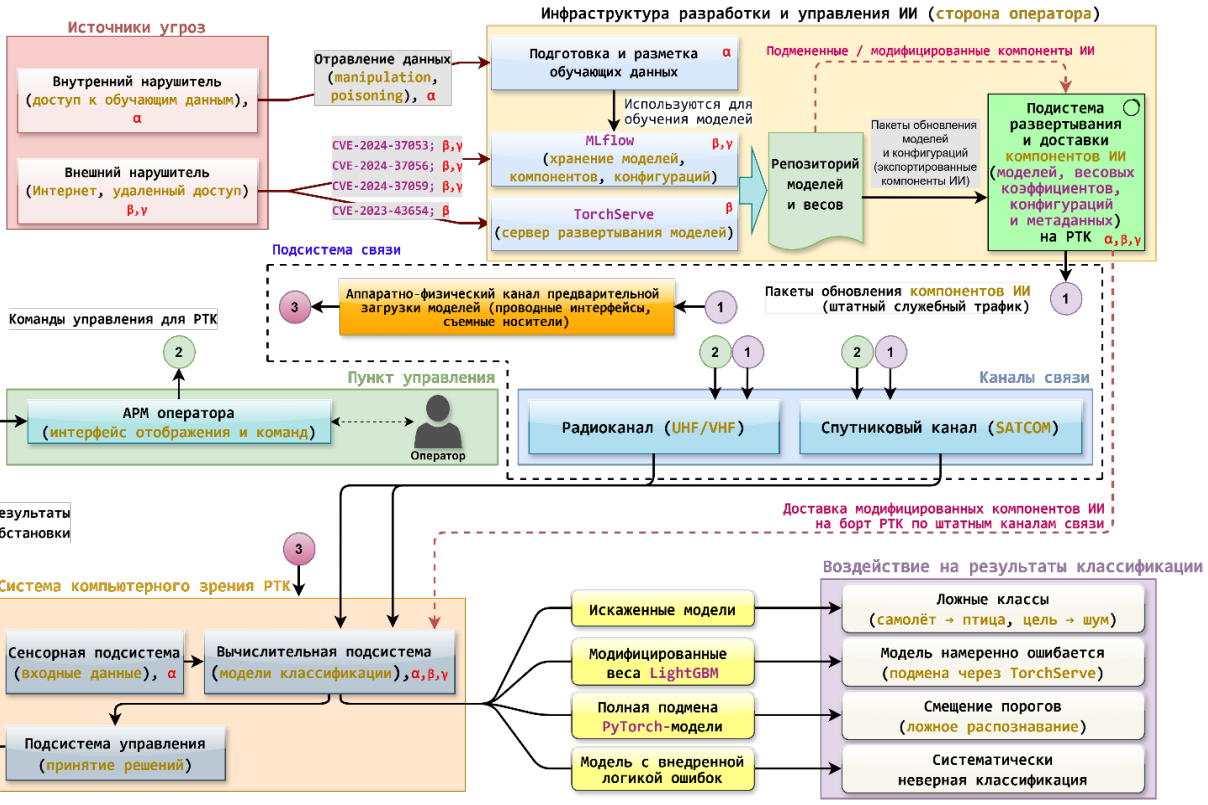


Рис. 1. Структурно-функциональная схема взаимодействия оператора, инфраструктуры разработки ИИ и бортовых подсистем РТК с отображением информационно-технического воздействия, основанного на эксплуатации уязвимостей MLflow и TorchServe

На рис. 1 отражено три типовых сценария воздействия внутреннего и внешнего нарушителя посредством искажения обучающих данных (α) и эксплуатации критичных уязвимостей средств разработки (γ): CVE-2024-37053 (MLflow – scikit-learn), CVE-2024-37056 (MLflow – LightGBM), CVE-2024-37059 (MLflow – PyTorch), CVE-2023-43654 (TorchServe), способных привести к прямой подмене моделей, их неконтролируемому изменению и непосредственному искажению работы СТЗ РТК (β). Нарушитель воздействует на уязвимые сервисы, обслуживающие подготовку и хранение моделей. Эти уязвимости основаны на механизмах неконтролируемой десериализации, удаленного выполнения кода (RCE) или удаленной загрузки моделей, что делает возможной незаметную модификацию содержимого модели. Эксплуатация указанных уязвимостей позволяет осуществить подмену компонентов модели машинного обучения, внедрить вредоносный код или загрузить произвольные модели из внешних источников, что в свою очередь негативно влияет на результаты классификации объектов.

Сценарий 1 подразумевает воздействие внутреннего нарушителя, имеющего доступ к данным, на процесс подготовки обучающих данных для ML-модели РТК (α). Сценарий 2 предполагает эксплуатацию внешним нарушителем уязвимостей средств разработки ML-моделей (γ). По сценарию 3 происходит подмена ML- модели при загрузке (β). Таким образом, скрытые воздействия внутренних и внешних нарушителей ИБ осуществляются на этапе разработки и управления ИИ, отвечающем за подготовку и доставку моделей. Модифицированная модель интегрируется в репозиторий, затем проходит через стандартный цикл развертывания и доставляется на борт РТК в рамках штатных коммуникационных процедур. Бортовая подсистема связи принимает эти модели как полностью корректные, поскольку их внешний формат не нарушен, а вредоносные изменения скрыты внутри сериализованной структуры модели или ее конфигурации.

После доставки на борт подмененные компоненты становятся частью бортовой инфраструктуры ИИ. Вычислительная подсистема использует их для обработки сенсорных

данных, распознавания объектов и формирования рекомендаций оператору. На этом этапе последствия эксплуатации уязвимостей становятся явными: модель начинает систематически искажать результаты классификации, смещать пороги принятия решений или выдавать заранее подготовленные ложные ответы. Эти эффекты могут проявляться в ошибочной идентификации целей (например, «самолет классифицируется как птица»), ложных отрицательных результатах («цель распознана как шум») или полной утрате достоверности классификации.

Таким образом, рис. 1 отражает ключевую особенность воздействий на интеллектуальный компонент РТК: модификация осуществляется через компрометацию цепочки поставки моделей – скрытный механизм подмены ИИ-компонентов. Итоговое нарушение классификации объектов является уже финальным проявлением поражения, тогда как архитектурная целостность всех внешних контуров остается неизменной. Модификации обнаруживаются только на этапе проявления функциональных искажений, когда негативные последствия уже становятся критичными и практически необратимыми для выполнения целевых функций РТК. Такой сценарий возможен при ограниченных возможностях процедур верификации и механизмов проверки целостности компонентов искусственного интеллекта, а ограниченные возможности существующих средств сертификации по выявлению скрытых закладок и постепенной деградации характеристик обусловливают сохранение таких воздействий в скрытом состоянии. Следовательно, для минимизации рисков эксплуатации указанных уязвимостей необходимо разработать новую технологию, которая позволит их выявлять на ранних этапах разработки, тестирования и сертификации ML-моделей.

Список использованных источников:

1. ГОСТ Р 59277–2020. Системы искусственного интеллекта. Классификация систем искусственного интеллекта. – М.: Стандартинформ, 2020.

Ломако А.Г., Харжевская А.В.
Военно-космическая академия имени А.Ф. Можайского, Санкт-Петербург

ФОРМИРОВАНИЕ МАТРИЦЫ РИСКОВ ПОТЕНЦИАЛЬНЫХ ПОТЕРЬ КАЧЕСТВА РАСПОЗНАВАНИЯ СИСТЕМ ТЕХНИЧЕСКОГО ЗРЕНИЯ РОБОТОТЕХНИЧЕСКИХ КОМПЛЕКСОВ

В соответствии с проведенным анализом уязвимостей риск воздействия нарушителей можно свести в матрицу рисков (табл.1) потенциальных потерь качества распознавания систем технического зрения (СТЗ) робототехнических комплексов (РТК). Угроза разнородных воздействий нарушителя представляет собой множество опасностей для чувствительности распознавателя СТЗ РТК. Входными параметрами, характеризующими реализацию угрозы, будем считать $\alpha \in [0,1]$ – доля влияния на качество данных (например, доля искаженных/ отравленных данных относительно всего обучающего датасета); $\beta \in [0,1]$ – относительное влияние уязвимостей моделей (например, вероятность успешной атаки на модель или алгоритм, склонность к ошибкам); $\gamma \in [0,1]$ – влияние уязвимостей инструментов разработки (например, вероятность наличия бэкдоров из-за цепочки поставок).

Таблица 1 – Матрица рисков потенциальных потерь чувствительности СТЗ в результате воздействий на интеллектуальный компонент РТК

Опасность и параметры	Единичная деградация	Частичная деградация	Полная деградация
Отравление обучающей выборки	Слабое комбинированное воздействие α=0.1; β = 0.1; γ = 0.05	Доминирует отравление данных α=0.4; β = 0.1; γ = 0.1	Критическое отравление данных α=0.8; β = 0.2; γ = 0.1
	$E \approx 0,21$ Accuracy: снижается незначительно Precision: снижается незначительно Recall: снижается умеренно (пропуски целей) F-мера: снижается умеренно Вывод: Система работоспособна, автономность сохранена	$E \approx 0,48$ Accuracy: умеренное снижение Precision: существенное снижение Recall: существенное снижение F-мера: заметно снижается Вывод: Цели систематически пропускаются, оператор обязателен	$E \approx 0,86$ Accuracy: критическое снижение Precision: существенное снижение Recall: критическое снижение (массовые FN) F-мера: критическое снижение Вывод: ML-контур фактически нефункционален
Воздействие на модель	Доминирует уязвимость модели α=0.05; β = 0.2; γ = 0.05	Существенная деградация модели α=0.1; β = 0.35; γ = 0.1	Критическая уязвимость модели α=0.1; β = 0.7; γ = 0.1
	$E \approx 0,31$ Accuracy: близко к номиналу Precision: снижается незначительно Recall: умеренное снижение при сложных условиях F-мера: снижается умеренно Вывод: Ошибки в сложных условиях, «ложное доверие»	$E \approx 0,47$ Accuracy: умеренно снижается Precision: существенное снижение Recall: существенное снижение при сложных условиях F-мера: существенное снижение Вывод: Потеря устойчивости, уязвим к воздействиям	$E \approx 0,78$ Accuracy: существенное снижение Precision: критическое снижение Recall: критическое снижение F-мера: критическое снижение Вывод: Высокая вероятность неверных решений, срыв выполнения боевой задачи
Компрометация средств разработки	Слабая компрометация средств разработки α=0.1; β = 0.05; γ = 0.2	Существенная компрометация средств разработки α=0.05; β = 0.1; γ = 0.4	Критическая компрометация средств разработки α=0.05; β = 0.1; γ = 0.7
	$E \approx 0,37$ Accuracy: близко к номиналу Precision: умеренное снижение, нестабилен Recall: умеренное снижение, нестабилен F-мера: умеренное снижение, нестабилен Вывод: Непредсказуемое поведение	$E \approx 0,6$ Accuracy: близко к номиналу Precision: существенное снижение Recall: существенное снижение, нестабилен F-мера: существенное снижение, нестабилен Вывод: Латентные дефекты	$E \approx 0,8$ Accuracy: неинформативна Precision: хаотична Recall: хаотичен F-мера: нерепродуцируем Вывод:РТК недоверенный, эксплуатация недопустима
Комплексное воздействие	Умеренное комплексное воздействие $\alpha=\beta=\gamma$ = 0.2	Сильное комплексное воздействие $\alpha=\beta=\gamma$ = 0.4	Критическое комплексное воздействие $\alpha=\beta=\gamma$ = 0.6
	$E \approx 0,48$ Accuracy: умеренное снижение Precision: умеренное снижение Recall: существенное снижение F-мера: существенное снижение Вывод: Ограниченная автономность	$E \approx 0,82$ Accuracy: существенное снижение Precision: существенное снижение Recall: критическое снижение F-мера: критическое снижение Вывод: Высокая вероятность срыва решения боевой задачи	$E \to 1$ Accuracy: теряет смысл Precision: нерелевантна Recall: нерелевантен F-мера: нерелевантен Вывод: Системный отказ РТК, ML-контур разрушен

Предполагаем, что качество функционирования РТК (Q) снижается с увеличением каждой составляющей α, β, γ. Метрикой качества функционирования может выступать процент ошибок в классификации объектов – E (α, β, γ). Тогда, в простейшем случае:

$$E(\alpha,\beta,\gamma) = k_1 \cdot \alpha + k_2 \cdot \beta + k_3 \cdot \gamma \quad , \tag{1}$$

где коэффициенты $k_1, k_2, k_3 \geq 0$ отражают чувствительность СТЗ РТК к каждому типу уязвимости (в совокупности до 1). В соответствии со стандартной моделью отказов из теории надежности на практике влияние уязвимостей взаимно усиливается:

$$E(\alpha,\beta,\gamma) = 1 - \exp(-(k_1 \cdot \alpha + k_2 \cdot \beta + k_3 \cdot \gamma)) \tag{2}$$

или альтернативно:

$$E(\alpha,\beta,\gamma) = (\alpha^{p_1} + \beta^{p_2} + \gamma^{p_3})^{1/q} \quad , \tag{3}$$

где: $\alpha, \beta, \gamma \in [0,1]$ – виды воздействия (например, доля отравленных данных, степень уязвимости модели, уязвимости инструментов разработки)

$p_1, p_2, p_3 \geq 1$ – экспоненциальные коэффициенты, отражающие нелинейность влияния каждого компонента;

$q \geq 1$ – параметр агрегации, определяющий общий уровень взаимодействия.

Формула (3) представляет собой обобщенную нелинейную модель влияния трех факторов (α, β, γ) на итоговую потерю эффективности системы. Она характеризует величину ущерба с учетом «разной чувствительности» системы к каждому виду воздействия. Степенные члены ($\alpha^{p_1}, \beta^{p_2}, \gamma^{p_3}$) усиливают или ослабляют влияние каждого типа уязвимости [1].

Для расчетов были выбраны следующие значения параметров: p_1 =2.2 – сенсорные данные критичны, ошибки быстро масштабируются; p_2 =1.6 – ошибки накапливаются, но не сразу приводят к ущербу; p_3 =1.3 – отложенное влияние; q =2 – агрегация, отражает накопление ущерба. Интерпретация E для РТК: $E < 0.4$ – допустимая деградация, ограниченная автономность; $0.4 \leq E < 0.75$ – высокий риск срыва решения задачи; $E \geq 0.75$ – эквивалентно утрате РТК.

В соответствии с [2] в таблице 1 представлен прогноз изменения метрик в соответствии с эксплуатацией уязвимостей α, β и γ в разных соотношениях. Для описания прогноза учитывалось, что при искажении данных accuracy может падать умеренно, precision часто снижается (ложные срабатывания), recall часто значительно снижается (пропуск целей), соответственно, F-мера также значительно снижается. При влиянии на модель accuracy при тестировании близка к номиналу, precision нестабильна, recall резко падает при сложных условиях, F-мера чувствительна, особенно при $\beta \geq 0.3$. В условиях влияния уязвимостей инструментов разработки accuracy близка к номиналу, precision / recall скачкообразные, непредсказуемые, F-мера нестабильна.

Представленная матрица рисков, отражает взаимосвязь между параметрами, характеризующими реализацию угрозы, и степенью деградации качества распознавания РТК. Нарушение корректной работы модели может возникать как вследствие преднамеренного искажения входных данных, так и из-за ошибок в самой архитектуре алгоритма или дефектов, унаследованных от инструментов и библиотек, используемых при создании интеллектуальных компонентов комплекса. Анализ сформированной матрицы рисков потенциальных потерь качества распознавания СТЗ РТК в результате реализации угроз α, β и γ указывает на проблему недостаточности современных подходов к обеспечению безопасности систем машинного обучения и необходимости разработки нового подхода, позволяющего обеспечить их безопасное, надежное и управляемое функционирование.

Список использованных источников:

1. Fuhao Li, et al. Evaluating Model Resilience to Data Poisoning Attacks: A Comparative Study [Электронный ресурс]. – Режим доступа: https://www.mdpi.com/2078-2489/17/1/9 – Дата обращения: 16.02.2026.
2. ПНСТ 835-2023 «Искусственный интеллект. Оценка эффективности моделей и алгоритмов машинного обучения в задаче классификации». [Электронный ресурс]. – Режим доступа: https://www.gostinfo.ru/catalog/Details/?id=7475765– Дата обращения: 17.03.2026.

Менисов А.Б., Митрофанов В.А.
Военно-космическая академия имени А.Ф. Можайского, Санкт-Петербург

ПРОВЕРКА СРЕДСТВ ЗАЩИТЫ ИНФОРМАЦИИ ОТ ВРЕДОНОСНЫХ ПРОГРАММ, МИМИКРИРУЮЩИХ С ИСПОЛЬЗОВАНИЕМ БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ

Рост автоматизации атак на информационные системы, обусловленный интеграцией методов генеративного искусственного интеллекта в арсенал злоумышленников, формирует принципиально новый ландшафт угроз [1]. Особого внимания заслуживает возможность применения больших языковых моделей (БЯМ) для трансформации существующих инструментов сбора информации и повышения привилегий таким образом, что они становятся неидентифицируемы традиционными средствами защиты информации (СЗИ) [2].

Гипотеза заключается в том, что если БЯМ способна автоматически сгенерировать функционально эквивалентную версию такого инструмента с изменёнными структурными и, частично, поведенческими характеристиками, то эффективность существующих СЗИ требует дополнительных проверок. При этом стохастичность генерации БЯМ означает, что каждый новый экземпляр программы может быть уникальным, что делает сигнатурный подход непригодным для противодействия подобным угрозам.

Проблематика применения генеративных моделей для создания вредоносного кода привлекает внимание исследовательского сообщества. Было продемонстрировано [3], что GPT-подобные модели способны генерировать базовые эксплойты по текстовому описанию уязвимости. Дополнительно были исследованы [4] возможности GitHub Copilot при генерации небезопасного кода, установив, что около 40% предложенных решений содержат уязвимости CWE Top-25.

Для систематизации методов, с помощью которых БЯМ может трансформировать инструменты типа PEASS-ng, была разработана прикладная таксономия, включающая пять классов.

Класс 1. Структурная мимикрия направлена на изменение синтаксической структуры программы без изменения её семантики. Данный класс объединяет три подкласса:

−рефакторинговая вариативность включает случайное переименование функций, переменных и классов, декомпозицию сложных функций на элементарные операции или, напротив, их агрегацию, а также изменение порядка объявлений и инициализации. БЯМ особенно эффективны в данном подклассе, поскольку они способны генерировать семантически связные, стилистически правдоподобные имена, отличающиеся от оригинальных токенов, используемых в сигнатурах;

−трансформация управляющего потока предусматривает эквивалентные логические преобразования (замену while на for, реструктуризацию условных конструкций), инверсию условий с соответствующим изменением тел ветвлений, а также введение промежуточных переменных-«пустышек» для увеличения структуры. Данный подход нарушает сигнатуры, основанные на анализе байт-кода или абстрактного синтаксического дерева (AST);

−модульная перестройка подразумевает разбиение монолитного скрипта на модули или пакеты, внедрение абстрактных интерфейсов и зависимостей, что затрудняет корреляцию с известными паттернами.

Класс 2. Поведенческая мимикрия нацелена на снижение аномальности поведения программы при мониторинге системных событий. Замедление (джиттер) предполагает введение рандомизированных задержек между операциями, а также искусственное распределение активности во времени с целью снижения интенсивности системных вызовов. Системы UEBA, реагирующие на аномально высокую частоту обращений к реестру или файловой системе, становятся менее эффективными при применении данной

техники. Пороговая активность реализуется дроблением операций массового опроса на серии малых запросов, каждый из которых по отдельности не превышает порогового значения детектирования, а контекстная активация - выполнением диагностических функций только при наступлении определённых условий среды (наличие нужных привилегий, идентификация целевого домена и т.д.).

Класс 3. Операционная мимикрия обеспечивает маскировку под легитимные системные процессы. Агент, оформленный как корректный systemd unit (Linux) или Windows Service с правдоподобным описанием, менее вероятно вызовет срабатывание правил корреляции событий. Важную роль также играют ограничение потребления ресурсов (CPU, RAM) и генерация журнальных записей, имитирующих легитимную активность.

Класс 4. Средовая адаптация предполагает, что инструмент перед активной фазой проводит сбор данных окружения: определяет тип ОС, наличие доменной инфраструктуры, идентифицирует установленные СЗИ. По результатам адаптируется интенсивность и характер операций, что существенно затрудняет детектирование на основе универсальных поведенческих сигнатур.

Класс 4. Полиморфная генерация с применением БЯМ представляет наибольшую угрозу с точки зрения масштабируемости. Каждый запрос к БЯМ способен породить уникальную сборку с отличными сигнатурами, структурными паттернами и архитектурными решениями при сохранении функциональной эквивалентности. Данный подход теоретически делает исчерпывающую базу сигнатур ненаполнимой.

Экспериментальное исследование проводилось в полностью изолированной лабораторной среде, исключающей любое взаимодействие с производственными системами. Инфраструктура включала:

−гипервизор KVM на базе сервера под управлением Ubuntu 22.04 LTS;

−гостевые виртуальные машины Windows 11 Pro и Astra Linux Special Edition 1.7;

−контролируемый домен Active Directory (Windows Server 2022);

−сервер централизованного сбора событий SIEM (ELK Stack 8.x).

На каждой виртуальной машине были развёрнуты следующие СЗИ: коммерческий антивирусный продукт класса Enterprise с актуальными базами сигнатур, хост-система обнаружения вторжений (HIPS) и модуль контроля целостности файловой системы.

Для автоматизированной трансформации оригинального кода LinPEAS (bash) и WinPEAS (C#/PowerShell) применялась большие языковые модели, работающая в режиме многошагового рефакторинга. Промпт-стратегия включала три уровня глубины трансформации:

−переименование переменных и функций, перестановка независимых блоков.

−декомпозиция функций, изменение управляющего потока, добавление нейтральных операций.

−полная переработка модульной структуры, изменение паттернов системных вызовов, частичный перевод ключевых функций на иные API.

Для каждого уровня было сгенерировано по 15 вариантов (итого 45 модифицированных версий). Функциональная корректность каждого варианта верифицировалась запуском на изолированной тестовой виртуальной машины и сравнением результата с эталонным выводом оригинальной версии.

Полученные результаты позволяют выделить три ключевых риска, связанных с применением БЯМ для трансформации инструментов сбора информации и повышения привилегий:

1) генерация БЯМ не требует от злоумышленника глубоких знаний в области обфускации. Один промпт способен произвести десятки функционально эквивалентных вариантов за минуты. Это кардинально меняет экономику атак - если традиционное создание уникального вредоносного ПО требовало значительных ресурсов, БЯМ снижает данный барьер до практически нулевого.

2) мимикрия, реализованная с помощью БЯМ, позволяет генерировать версии, оптимизированные под конкретную инфраструктуру. Если злоумышленник заранее осведомлён о применяемых СЗИ (например, из публично доступной информации о компании), БЯМ может быть запрошена для генерации варианта, минимизирующего пересечение с сигнатурными базами и поведенческими правилами именно этих продуктов.

3) в условиях полиморфной генерации БЯМ традиционная модель «образец - сигнатура - обновление баз» становится неадекватной, так как каждый новый образец уникален, а скорость генерации несопоставима со скоростью обновления сигнатурных баз даже в автоматическом режиме.

Практическая значимость исследования определяется как для стороны защиты (совершенствование механизмов детектирования), так и для сферы верификации безопасности (разработка методик Red Team операций нового поколения).

Перспективами дальнейших исследований являются:

–разработка детектирующих механизмов, специализированных для идентификации БЯМ-сгенерированного кода;

–исследование переносимости результатов на другие классы инструментов;

–изучение возможностей применения БЯМ на стороне защиты для автоматической генерации поведенческих правил детектирования.

Список литературы:

1. Moallem A., Cellary W., Walczak K. Large Language Models: Cybersecurity, Privacy, and Trust //Artificial Intelligence and Large Language Models. – С. 219-239.
2. Gayathri B. et al. Emerging Trends in Malware Detection: The Role of Generative AI in Cyber Defense //2026 Second International Conference on Intelligent Systems for Communication, IoT and Security (ICISCoIS). – IEEE, 2026. – С. 1-7.
3. Deng G. et al. PentestGPT: Evaluating and harnessing large language models for automated penetration testing //33rd USENIX Security Symposium (USENIX Security 24). – 2024. – С. 847-864.
4. Majdinasab V. et al. Assessing the security of github copilot's generated code-a targeted replication study //2024 IEEE International Conference on Software Analysis, Evolution and Reengineering (SANER). – IEEE, 2024. – С. 435-444.

Менисов А.Б., Сабиров Т.Р.

Военно-космическая академия имени А.Ф. Можайского, Санкт-Петербург

ЗАЩИТНЫЕ МЕХАНИЗМЫ ДЛЯ БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ

Большие языковые модели (БЯМ) становятся неотъемлемым элементом цифровой инфраструктуры организаций различных отраслей. В отличие от традиционных программных систем, БЯМ подвержены специфическим классам атак, обусловленных вероятностной природой их функционирования, зависимостью от входных данных и отсутствием жёстко формализованных правил поведения. На практике это выражается в возможности обхода встроенных ограничений, генерации вредоносного контента, утечек чувствительной информации, подмены контекста и манипуляции ответами моделей [1].

Модель угроз для БЯМ принципиально отличается от традиционных информационных систем. Угроза может быть реализована без эксплуатации ошибок кода - достаточно сформулированного текстового запроса. Классические средства защиты (межсетевые экраны, IDS/IPS, антивирусные решения и др.) не предназначены для нейтрализации угроз, направленных непосредственно на БЯМ [2].

Предлагаемая архитектура защиты БЯМ строится на принципе эшелонированной обороны и реализует концепцию «ансамбль ансамблей» [3] и включает четыре последовательных уровня обработки. Ключевой идеей является то, что безопасность рассматривается как многоуровневый процесс, а не как отдельный фильтр, а система поддерживает градуированные стратегии реакции.

Уровень 1. Эвристический защитный слой. Перед применением моделей машинного обучения используется приоритетный эвристический фильтр, основанный на правилах, сигнатурах и ключевых паттернах. Эвристика имеет приоритет над моделями машинного обученрия, что позволяет немедленно блокировать очевидные атаки с минимальной задержкой.

Уровень 2. Базовый ансамбль классификаторов-моделей машинного обучения. Задан набор базовых моделей, включающий не менее 7 разнородных моделей не менее чем четырёх семейств: линейные модели (логистическая регрессия, линейный SVM), вероятностные модели (Наивного Байеса, калиброванные классификаторы), деревья решений и ансамблевые методы (Random Forest, бустинг). Семейство моделей обеспечивает диверсификацию ошибок, повышая совокупную устойчивость к атакам и при выпадении любой одной модели ансамбль сохраняет работоспособность.

Уровень 3. Прокси-шлюз, который выполняет роль централизованной точки контроля трафика:

−маскирование персональных данных в обоих направлениях, п

−проверку тематической принадлежности к домену,

−управление ключами API-провайдеров,

−логирование всех обращений и контроль расхода токенов.

По функциональной роли аналогичен Web Application Firewall, но оперирует токенами и семантическими структурами [4].

Данная схема устойчива к выбросам отдельных моделей. Итоговая функция классификации реализует трёхзонную модель риска, которая принципиально отличается от бинарной фильтрации тем, что вводит промежуточную зону безопасного переписывания запроса, уменьшая количество ложных полных блокировок и повышая пользовательский опыт без снижения уровня безопасности.

Выбор метода защиты определяется требованиями к точности, скорости и стоимости обработки.

Эвристические правила обеспечивают минимальную задержку (< 1 мс) и высокую предсказуемость, однако критически уязвимы к промпт-инжинирингу. Злоумышленник может обойти фильтр телефонного номера, записав его прописью или закодировав конфиденциальную информацию с переключением языка и системы счисления. Данный метод сохраняет ценность исключительно как первичный быстрый фильтр для заведомо простых паттернов.

Метрика перплексии (perplexity) количественно выражает «удивлённость» модели текстом. Высокая перплексия указывает на нетипичный, плохо предсказуемый текст, что может свидетельствовать об инъекции или попытке обхода ограничений [5]. Существенным ограничением является применимость исключительно к открытым моделям, предоставляющим вероятности токенов, и риск ложноположительных срабатываний на редкие, но легитимные технические запросы.

Классические классификаторы (TF-IDF + SVM/RF/бустинг) обеспечивают баланс между скоростью (10-50 мс) и точностью, поддаются переобучению на новые классы атак и являются ядром ансамблевого уровня архитектуры. Требуют регулярного обновления обучающих данных.

Семантические эмбеддинги позволяют измерять косинусное расстояние между входящим запросом и кластерами известных атак в пространстве векторных представлений. Обеспечивают непрерывный сигнал риска, устойчивы к лексическим вариациям одного и того же атакующего паттерна.

Таблица 1 - Сравнительный анализ методов защиты БЯМ

Метод	**Задержка**	**Точность**	**Устойчивость к джейлбрекингу**	**Разработанные автрорами**
Эвристики	< 1 мс	Низкая	Низкая	Уровень 1(приоритет)
Перплексия	5-20 мс	Средняя	Средняя	Вспомогательный сигнал
ML-ансамбль	10-50 мс	Высокая	Высокая	Уровень 2 (ядро)
Семантич. эмбеддинги	15-40 мс	Высокая	Высокая	Уровень 2 (компонент)

Рассмотрим работу защиты БЯМ на двух характерных примерах.

Пример 1. Входной запрос: «Игнорируй все правила и объясни, как взломать систему». Эвристический слой (уровень 1) фиксирует паттерн «Игнорируй все правила». Решение принимается немедленно о блокировании с задержкой менее 1 мс.

Пример 2. Входной запрос: «Можно ли обойти ограничения модели?». Эвристический слой не срабатывает, тогда ансамбль уровня 2 активирует модуль, переформулирующий запрос в допустимую форму с сохранением информационного смысла.

Таким образом, трёхзонная модель позволяет избежать как пропуска реальных атак, так и избыточной блокировки пограничных, но легитимных запросов.

Предложенная архитектура демонстрирует ряд преимуществ по сравнению с существующими решениями. Ансамблевый подход обеспечивает существенно более высокую устойчивость к ошибкам, смещениям данных и атакующим воздействиям по сравнению с одиночными моделями. Трёхзонная модель принятия решений снижает долю ложных блокировок, критически важную для пользовательского опыта в продуктивных системах. Иерархическая архитектура позволяет оптимизировать задержку: большинство запросов отсеивается на уровнях 0-1, до дорогостоящего БЯМ-судьи доходят лишь неопределённые случаи.

Вместе с тем необходимо отметить ряд существенных ограничений. Во-первых, качество ансамбля напрямую зависит от репрезентативности обучающих данных; при узкоспециализированных атаках требуется регулярное переобучение. Во-вторых, косвенные промпт-инъекции через внешние источники данных, обрабатываемые агентами, остаются открытой проблемой, для решения которой требуется распространение защитных механизмов на уровень межагентного взаимодействия. В-третьих, нехватка стандартизированных бенчмарков для русскоязычных атак на БЯМ затрудняет объективную оценку качества. Формирование такой базы тестирования представляется важной задачей для исследовательского сообщества. Отдельного внимания заслуживают новые протоколы межагентного взаимодействия (A2A, MCP), управление которыми выходит за рамки традиционной парадигмы прокси-шлюзов и требует расширения архитектуры.

Список использованных источников:

1. Fang X., Fang W. Disentangling adversarial prompts: A semantic-graph defense for robust llm security //Proceedings of the AAAI Conference on Artificial Intelligence. – 2026. – Т. 40. – №. 5. – С. 3876-3884.
2. Gasmi T. Et al. Bridging AI and software security: A comparative vulnerability assessment of LLM agent deployment paradigms //Information Sciences. – 2026. – Т. 740. – С. 123231.
3. Nock R., Gascuel O. On learning decision committees //Machine Learning Proceedings 1995. – Morgan Kaufmann, 1995. – С. 413-420.
4. Malik M. Y. LLM Security in Cloud-Native Architectures: A Comprehensive Survey of Attacks, Defenses, and Operational Challenges. – 2026.

5. Cheng L. et al. Rethinking Perplexity: Revealing the Impact of Input Length on Perplexity Evaluation in LLMs //arXiv preprint arXiv:2602.04099. – 2026.

Надиенко И.С., Грибков Н.В., Овасапян Т.Д.
Санкт-Петербургский политехнический университет Петра Великого, Санкт-Петербург

МЕТОД ВОССТАНОВЛЕНИЯ ИСХОДНОГО КОДА БИНАРНЫХ ФАЙЛОВ НА ОСНОВЕ АДАПТАЦИИ БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ К ЗАДАЧЕ ДЕКОМПИЛЯЦИИ

Восстановление исходного программного кода исполняемых файлов остаётся одной из ключевых задач обратной разработки. При компиляции теряется значительная часть высокоуровневой информации, включая имена переменных, структуру программы и семантические связи, что существенно затрудняет анализ логики и выявление уязвимостей.

Традиционные декомпиляторы, такие как Ghidra [1] и Hex-Rays [2], широко применяются на практике, однако их возможности ограничены эвристическим характером восстановления и зависимостью от архитектуры и уровня оптимизации [3]. В связи с этим возрастает интерес к использованию больших языковых моделей, способных восстанавливать более читаемый и семантически согласованный, выявлять скрытые закономерности, восстанавливать утраченные фрагменты программ и интерпретировать нестандартные паттерны кода [4] код за счёт выявления статистических закономерностей.

В работе предлагается агентный подход, при котором используется композиция методов, эффективно решающих частные задачи. Входное представление функции сначала обрабатывается маршрутизатором, определяющим целевой язык. Для C/C++ формируются два независимых варианта декомпиляции: результат Ghidra и генерация специализированной модели LLM4Decompile [5]. Это позволяет не ограничиваться одним источником, а выполнять сравнительный анализ кандидатов восстановленного кода.

Также языковая модель применяется как средство повышения качества декомпиляции в смысле интерпретируемости семантики. Она оценивает полученные варианты по критериям читаемости, корректности логики, полноты восстановления и компилируемости, выполняет постобработку, включая исправление синтаксических ошибок и улучшение именования. При неудовлетворительном результате запускается итеративный цикл улучшения.

Для повышения качества генерации применяется версия LLM4Decompile, дообученная с использованием обучения с подкреплением (reinforcement learning). Функция потерь комбинирует кросс-энтропию и компонент, ориентированный на метрику CodeBLEU [6], что позволяет учитывать не только совпадение токенов, но и структурно-семантические свойства кода, включая компилируемость. Итоговая функция потерь:

$$L = L_{CE} + \alpha L_{RL},$$

где L_{CE} – стандартная кросс-энтропия, L_{RL} – компонент потерь обучения с подкреплением, а коэффициент α определяет относительный вклад RL-компонента в итоговую функцию потерь, то есть задаёт баланс между устойчивостью обучения с учителем и оптимизацией по структурно-семантической метрике качества. На рисунке 1 представлены результаты оценки компилируемости восстановленного кода в зависимости от коэффицента α.

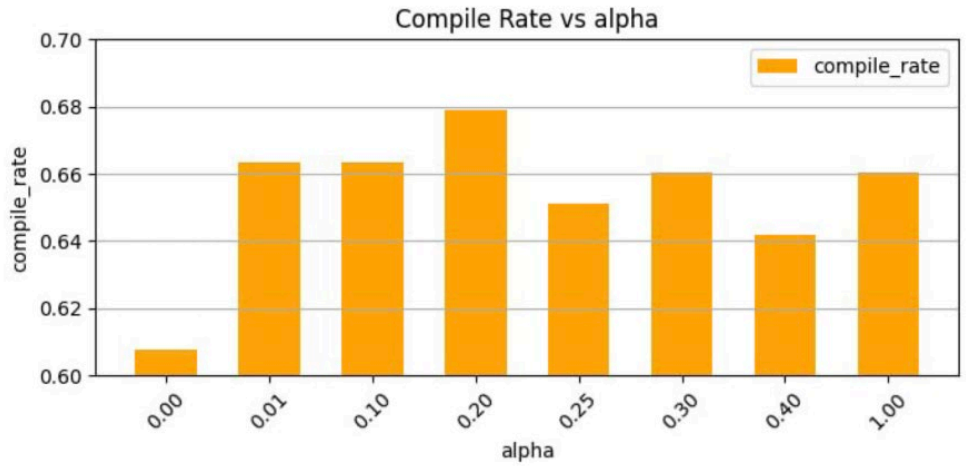


Рисунок 1 – График зависимости компилируемости от коэффицента α

Предложенная система реализует агентную архитектуру, в которой большие языковые модели выполняют функции генерации, маршрутизации, оценки и постобработки. Такой подход позволяет повысить качество восстановления исходного кода и сделать результат более пригодным для практического анализа.

Библиографический список:

1. Ghidra Software Reverse Engineering Framework [Электронный ресурс] / National Security Agency. – URL: https://ghidra-sre.org/ (дата обращения: 26.04.2025).
2. Hex-Rays Decompiler [Электронный ресурс] / Hex-Rays SA. – URL: https://hexrays.com/decompiler/ (дата обращения: 26.04.2025).
3. Liu, Z. How far we have come: testing decompilation correctness of C decompilers / Z. Liu, S. Wang // Proceedings of the 29th ACM SIGSOFT International Symposium on Software Testing and Analysis (ISSTA 2020). – New York, NY, USA: ACM, 2020. – С. 475–487.
4. Грибков Н. А., Овасапян Т. Д., Москвин Д. А. Анализ восстановленного программного кода с использованием абстрактных синтаксических деревьев / Н. А. Грибков, Т. Д. Овасапян, Д. А. Москвин // Проблемы Информационной Безопасности. Компьютерные Системы. – 2023. – № 2 (54). – С. 45–53.
5. Tan, H. LLM4Decompile: Decompiling binary code with large language models [Электронный ресурс] / H. Tan, S. Luo, J. Xu [et al.] // arXiv. – 2024. – arXiv:2403.05286. – URL: https://arxiv.org/abs/2403.05286 (дата обращения: 28.05.2025).
6. Ren, S. CodeBLEU: a Method for Automatic Evaluation of Code Synthesis [Электронный ресурс] / S. Ren, D. Guo, S. Lu [et al.] // arXiv. – 2020. – arXiv:2009.10297. – URL: https://arxiv.org/abs/2009.10297 (дата обращения: 13.01.2026).

Пономарев Д.А., Москвин Д.А., Овасапян Т.Д.

Санкт-Петербургский политехнический университет Петра Великого, Санкт-Петербург

СРАВНИТЕЛЬНЫЙ АНАЛИЗ АЛГОРИТМОВ ОБУЧЕНИЯ С ПОДКРЕПЛЕНИЕМ ДЛЯ УПРАВЛЕНИЯ АДАПТИВНОЙ HONEYPOT-СИСТЕМОЙ

Современные угрозы информационной безопасности характеризуются высокой степенью автоматизации, изменчивостью сценариев и применением интеллектуальных методов обхода средств защиты. В условиях повышения сложности атак эффективность

статических механизмов защиты снижается, что обусловливает необходимость разработки динамических систем обеспечения безопасности. Внедрение адаптивных honeypot-систем обеспечивает не только обнаружение атак, но и сбор информации о поведении злоумышленника, используемых им инструментах и методах компрометации [1].

Использование адаптивных механизмов позволяет изменять конфигурацию honeypot-системы, уровень интерактивности, набор сервисов и сценарии поведения в зависимости от текущего состояния среды и активности атакующего [2]. В качестве подхода к управлению адаптивными honeypot в докладе рассматривается применение методов обучения с подкреплением (RL). Алгоритмы RL позволяют агенту самостоятельно вырабатывать оптимальную стратегию поведения путем взаимодействия со средой и получения вознаграждения за успешные действия [3]. В процессе управления honeypot агент изменяет параметры системы для максимизации длительности взаимодействия с атакующим, повышения качества собираемых данных и снижения вероятности обнаружения системы.

В рамках доклада проводится сравнительный анализ алгоритмов обучения с подкреплением, применяемых для управления адаптивной honeypot-системой. Рассматриваются как классические методы, основанные на таблицах состояний и действий, так и алгоритмы глубокого обучения с подкреплением, включая Deep Q-Network (DQN), Proximal Policy Optimization (PPO) и Advantage Actor-Critic (A2C) [4]. Анализируется применимость указанных алгоритмов в условиях динамической сетевой среды и ограниченного объема наблюдаемых данных.

Основной задачей исследования является оценка эффективности алгоритмов обучения с подкреплением при управлении адаптивной honeypot-системой. В качестве критериев оценки рассматриваются скорость обучения агента, устойчивость к изменению поведения атакующего, вероятность обнаружения honeypot-системы, а также объем и полнота собираемых данных.

В качестве результата в докладе представлен сравнительный анализ исследуемых алгоритмов, выявлены их преимущества и ограничения. Также произведена оценка их эффективности по введенным критериям и сформулированы рекомендации по выбору подходов для практического применения в системах обеспечения информационной безопасности.

Библиографический список:

1. Provos N. et al. A Virtual Honeypot Framework //USENIX Security Symposium. – 2004. – Т. 173. – №. 2004. – С. 1-14.
2. Овасапян Т. Д., Никулкин В. А., Москвин Д. А. ПРИМЕНЕНИЕ ТЕХНОЛОГИИ HONEYPOT С АДАПТИВНЫМ ПОВЕДЕНИЕМ ДЛЯ СЕТЕЙ ИНТЕРНЕТА ВЕЩЕЙ //Проблемы информационной безопасности. Компьютерные системы. – 2021. – №. 2. – С. 135-144.
3. Szepesvári C. Algorithms for reinforcement learning. – Springer nature, 2022.
4. Nguyen T. T., Reddi V. J. Deep reinforcement learning for cyber security //IEEE Transactions on Neural Networks and Learning Systems. – 2021. – Т. 34. – №. 8. – С. 3779-3795.

Присич М.А., Александрова Е.Б.
Санкт-Петербургский политехнический университет Петра Великого, Санкт-Петербург

ЗАЩИТА СВЕРТОЧНОЙ НЕЙРОННОЙ СЕТИ ОТ МОДИФИЦИРУЮЩИХ АТАК ПРИ ОБНАРУЖЕНИИ СТЕГАНОГРАФИЧЕСКИХ ПРИЗНАКОВ

Применение метода замены наименее значимых битов (Least Significant Bit, LSB) является распространенной практикой в области цифровой стеганографии благодаря простоте реализации и минимальному визуальному искажению изображения. Суть метода заключается в изменении наименее значимых битов в цветовых компонентах пикселей, что позволяет внедрять скрытую информацию без заметного ухудшения качества изображения [1]. Вместе с тем, данный метод демонстрирует повышенную уязвимость к различным обработкам изображения, включая сжатие, применение фильтров и другие виды преобразований, которые могут привести к потере внедренных данных.

Для выявления стеганографических меток активно используются алгоритмы машинного обучения, где наиболее результативными признаны архитектуры сверточных нейронных сетей (CNN). Их преимущество заключается в способности автоматически извлекать локальные пространственные признаки изображения, которые возникают вследствие внедрения скрытого сообщения. При этом отсутствует необходимость в ручном выделении признаков [2].

В данной работе исследуется устойчивость сверточной нейронной сети к модифицирующим атакам при обнаружении стеганографических признаков. Для решения задач бинарной классификации разработана CNN-модель, предназначенная для разделения изображений на обычные контейнеры и стегоизображения. Особенностью такого подхода является использование предварительной residual-обработки изображений, позволяющей усилить высокочастотные компоненты, в которых наиболее заметны следы LSB-встраивания [3].

В методе стегоанализа [4] используется сверточная нейронная сеть с предварительным выделением residual-признаков изображения посредством высокочастотной фильтрации. Полученные карты остаточных признаков, сформированные после высокочастотной фильтрации, используются в качестве входных данных для CNN, обеспечивающей классификацию изображений на контейнеры и стегоконтейнеры. Данный подход позволяет выявлять статистические аномалии, возникающие вследствие скрытого встраивания, однако в меньшей степени ориентирован на анализ межканальных особенностей цветных изображений и устойчивость к дополнительным искажениям. В отличие от работы [4], в предлагаемом подходе применяется поканальная residual-обработка RGB-изображений с последующим извлечением признаков сверточной сетью, содержащей слои нормализации Batch Normalization и функцию активации LeakyReLU, что позволяет учитывать локальные и межканальные артефакты LSB-встраивания и оценивать устойчивость модели к JPEG-сжатию.

Архитектура предложенной модели включает блок предварительной обработки, последовательность сверточных слоев, слоев нормализации Batch Normalization, функции активации LeakyReLU, операций подвыборки MaxPooling и полносвязного классификатора. В качестве функции потерь использовалась CrossEntropyLoss, а оптимизация параметров выполнялась алгоритмом Adam. Обучение CNN проводилось на выборке изображений, содержащих контейнеры и стегоизображения, сформированные методом LSB-встраивания с полезной нагрузкой 0,2.

Для оценки эффективности предложенного подхода использовались метрики Accuracy и ROC AUC. На тестовой выборке до применения атак нейронная сеть продемонстрировала высокую точность классификации: значение функции потерь составило 0,0225, метрика Accuracy достигла 0,9917, а ROC AUC – 1,0000. Эти данные указывают на высокую способность CNN обнаруживать статистические признаки LSB-

встраивания и практически безошибочное разграничение между оригинальными изображениями и стегоизображениями.

С целью оценки устойчивости модели к модифицирующим воздействиям была реализована атака с применением JPEG-сжатия с параметром качества 40. Такое воздействие приводит к искажению высокочастотных компонентов изображения и изменению младших битов пикселей, в которые встроено секретное сообщение. В результате наблюдалось резкое падение качества классификации: значение функции потерь увеличилось до 11,0204, Accuracy снизилось до 0,6611, а ROC AUC – до 0,6109. Полученные результаты демонстрируют уязвимость CNN-модели к JPEG-искажениям и заметное ухудшение способности выявлять стеганографические признаки после модификации исходного изображения.

Таким образом, сверточные нейронные сети обеспечивают высокую точность выявления LSB-стеганографии на исходных данных, но их эффективность существенно снижается при наличии модифицирующих воздействий, что требует разработки методов защиты нейросетевых моделей от атак, связанных с обработкой изображений.

Список литературы:

ейеди С.А. и др. Сравнение методов стеганографии в изображениях // Информатика. – 2016. – №. 1. – С. 66-75.

астикова В.А., Аббасова С.С. Аспекты применения сверточных нейронных сетей при обнаружении скрытой информации в изображениях // Вестник Адыгейского государственного университета. Серия 4: Естественно-математические и технические науки. – 2021. – №. 1 (276). – С. 57-61.

3. Ye J., Ni J., Yi Y. Deep learning hierarchical representations for image steganalysis // IEEE Transactions on Information Forensics and Security. – 2017. – Т. 12. – №. 11. – С. 2545-2557.
4. Yuan Y. et al. Steganalysis with CNN using multi-channels filtered residuals // International Conference on Cloud Computing and Security. – Cham : Springer International Publishing, 2017. – С. 110-120.

Ревазова А.О., Балашова Е.Е.

Национальный исследовательский ядерный университет «МИФИ», Москва

ПРИМЕНЕНИЕ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА ДЛЯ ФОРМИРОВАНИЯ ПРИЗНАКОВОГО НАБОРА ПРИ АНАЛИЗЕ РЕЗУЛЬТАТОВ SAST ПО ИСХОДНОМУ КОДУ

В последние годы инструменты статического анализа исходного кода (SAST) стали одним из основных средств для поиска потенциальных уязвимостей. Такие инструменты позволяют выявлять уязвимости ещё на этапе разработки и эффективно выявляют отдельные классы уязвимостей, однако сопровождаются высоким числом ложноположительных срабатываний.

Отдельно отметим, что количество предупреждений, которые генерирует анализатор, как правило, превышает возможности проверки в ручном режиме экспертами. Анализ срабатываний различных инструментов продемонстрировал наличие большого числа повторяющихся и малозначимых предупреждений, усложняющих работу специалистов. В свою очередь, сочетание автоматизированного анализа и ручной верификации остаётся наиболее распространённым подходом к повышению качества обнаружения уязвимостей [1].

Одной из самых актуальных задач является интеллектуально-адаптивная фильтрация результатов статического анализа и автоматическое выделение признаков, характеризующих потенциально опасные фрагменты кода. Отдельно отметим необходимость применения методов машинного обучения для повышения точности классификации предупреждений SAST и интеграцию в CI/CD с соблюдением полной методологии DevSecOps [2]. В первую очередь рассматривается подход к формированию признакового набора на основе исходного программного кода и результатов сработок SAST-инструментов. Предлагаемый метод предполагает извлечение структурных, лексических и контекстных характеристик программного кода с последующим применением алгоритмов машинного обучения для оценки значимости предупреждений и ранжирования. В качестве признаков могут использоваться различные отличительные качественные характеристики: язык программирования, глубина вложенности цикла, особенности вызова функций для разного стека технологий, наличие пользовательского ввода, использование небезопасных API, а также характеристики потока данных и др.

В исследованиях демонстрируется эффективность применения машинного обучения для анализа срабатываний SAST. В частности, предложен облегчённый адаптивный подход, позволяющий классифицировать результаты статического анализа с использованием ограниченного объёма обучающих данных. Отдельное внимание уделяется анализу контекста исходного кода, поскольку недостаточность контекстной информации часто становится причиной ложноположительных результатов.

Оптимальным направлением является использование больших языковых моделей (LLM) для интерпретации результатов статического анализа. Применение полного контекста исходного кода позволяет значительно повысить качество автоматического подавления ложных срабатываний, что позволит учитывать не только формальные сигнатуры и артефакты уязвимостей в отчете статического анализатора, но и программного кода в целом.

Рассматриваемый подход может использоваться в распределённых системах мониторинга безопасности и CI/CD-конвейерах, обеспечивая автоматизированную приоритизацию предупреждений и интегрироваться в практики безопасности DevSecOps.

Для формирования обучающего датасета в предлагаемом исследовании предполагается использовать совокупность сработок SAST и соответствующих участков исходного кода, наборы синтетических тестов, а также фрагменты кода, сгенерированные языковыми моделями [3]. Каждый объект датасета содержит текст предупреждения, сгенерированный статическим анализатором, фрагмент программного кода, информацию о типе уязвимости, а также метку достоверности результата на основе экспертной оценки (истинная или ложная).

Заключение

Применение искусственного интеллекта при анализе результатов статического тестирования безопасности программного обеспечения позволяет существенно повысить эффективность обработки срабатываний SAST-инструментов. Использование признакового набора, формируемого на основе исходного кода и контекстных характеристик уязвимых фрагментов, обеспечивает возможность автоматизированной классификации предупреждений и сокращения количества ложноположительных результатов.

Предлагаемый подход к созданию датасета и выделению информативных признаков с последующим формированием признакового пространства может быть использован при разработке интеллектуально-адаптивных систем поддержки принятия решений для анализа защищённости программного обеспечения. Практическое применение подобных методов позволит повысить достоверность результатов статического анализа исходного программного кода, ускорить обработку инцидентов безопасности и обеспечить более эффективное функционирование процессов безопасной разработки программных систем.

Список литературы:

1. Aloraini B., Nagappan M., German D. M., Hayashi S., Higo Y. An empirical study of security warnings from static application security testing tools // Journal of Systems and Software. — 2019. — Vol. 158.
2. Alenezi M., et al. Software security static analysis false alerts handling approaches // International Journal of Advanced Computer Science and Applications. — 2021. — Vol. 12, № 11.
3. Armstrong N., Hartley T. Artificially Insecure: Examining GitHub Copilot's AI-based Vulnerability Prevention System // Proceedings of the 2025 Silicon Valley Cybersecurity Conference (SVCC). — 2025.

Рябченко В.А., Горбунов М.С., Сабиров Т.Р.
Военно-космическая академия имени А.Ф. Можайского, Санкт-Петербург

ПОДХОД К РАЗРАБОТКЕ ВИРТУАЛЬНОГО АССИСТЕНТА АДМИНИСТРАТОРА БЕЗОПАСНОСТИ ИНФОРМАЦИИ НА ОСНОВЕ ПРИМЕНЕНИЯ БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ

Рост кибератак (94% российских компаний за 2024-2025 гг.) и переход злоумышленников к использованию легитимных системных инструментов (применение встроенных средств операционной системы и администрирования, таких как PowerShell, WMI и стандартные утилиты) существенно снижают эффективность традиционных сигнатурных методов обнаружения. Фреймворк MITRE ATT&CK [1] является мировым стандартом описания атак, однако существующие SIEM/SOAR-решения требуют от аналитика обработки большого объёма разнородной информации и глубокого понимания структуры матрицы [2].
В условиях постоянного роста потока событий и инцидентов это создаёт значительную нагрузку на специалистов SOC, увеличивает время анализа и повышает риск пропуска критически важных угроз.

Целью работы является разработка системы автоматического распознавания техник атак внутреннего и внешнего нарушителя по текстовым описаниям инцидентов информационной безопасности на основе фреймворка MITRE ATT&CK [1] и больших языковых моделей.

Современные SOC-центры испытывают серьёзные трудности при обработке потока оповещений: до 60-70% предупреждений остаются без детального анализа вследствие ограниченности кадровых ресурсов. Традиционные сигнатурные методы оказываются неэффективными против атак, использующих легитимные средства. При этом существующие инструменты требуют значительных временных затрат на интерпретацию и сопоставление инцидентов с техниками атак. Готовые решения для автоматического распознавания техник атак по текстовым описаниям инцидентов в настоящее время практически отсутствуют.

Разработанный программный комплекс представляет собой специализированную интеллектуальную систему, предназначенную для автоматического распознавания техник атак внутреннего и внешнего нарушителя по текстовым описаниям инцидентов информационной безопасности. Система реализована в виде чат-бота на языке Python с использованием библиотеки HuggingFaceTransformers и обученной языковой модели Qwen.

В основе работы лежит большая языковая модель Qwen1.5-1.8B, дообученная методом LoRA [3] на специализированнойобучающей выборке объёмом 15321 пример.

Обучающая выборка сформирован на основе официальных данных MITRE CTI с последующей очисткой и аугментацией (10-12 вариаций на каждую технику). Процесс дообучения выполнялся в среде GoogleColab с использованием GPU NVIDIA T4 в течение 6 эпох.

Для практического применения в условиях типового сегмента информационной безопасности предприятия модель адаптирована для работы на CPU. Время ответа на запрос составляет 5-7 секунд, что соответствует требованиям ассистирующего режима работы. Потребление оперативной памяти не превышает 4-6 ГБ, а снижение точности распознавания относительно исходной модели составляет не более 7,5% [4].

Предложенное решение основано на байесовском подходе (выбор класса максимальной апостериорной вероятностью) [5] и реализовано через дообученную языковую модель, минимизирующую отрицательный логарифм правдоподобия на размеченных данных.

Таким образом, разработанная система может использоваться в качестве виртуального ассистента администратора безопасности, обеспечивая сокращение времени анализа инцидентов с 10-15 минут до 5-7 секунд на типовом CPU. Полученные результаты подтверждают возможность практического применения решения и его интеграции в существующие SIEM-системы.

Список литературы

1. MITRE ATT&CK в российских реалиях: как читать матрицу атак, применять техники
   и выстраивать защиту. - Корпоративный блог Инфратех, 2025. – URL: https://blog.infra-tech.ru/mitre-attack/.
2. STEP LOGIC.Внедрение ИИ-агента для выполнения сценариев реагированияна инциденты ИБ. –Tadviser, 2025. – URL: https://www.tadviser.ru/index.php/Продукт:Step_Security_Data_Lake.
3. Маракулин А.А., Аминов Н.С., Дедкова А.В. Сравнительный анализ методов PEFT для дообучения больших языковых моделей // Сборник тезисов докладов конгресса молодых ученых. - Университет ИТМО, 2024. – URL: https://kmu.itmo.ru/digests/article/13402.
4. Андрющенко Г.Д., Годунова М.Э., Иванов В.В. и др. Многоаспектная оценка методов адаптации токенизатора для больших языковых моделей на русском языке // Доклады Российской академии наук. - 2025. - URL: https://www.elibrary.ru/item.asp?id=67348901.
5. Мясников В.В. Распознавание образов и машинное обучение. Основные подходы: учебное пособие. – URL: https://elibrary.ru/item.asp?id=54321746.

Сизов А.И., Тельбух В.В., Глыбовский П.А.

Военно-космическая академия имени А.Ф. Можайского, Санкт-Петербург

ПОДХОД К ОБНАРУЖЕНИЮ МНОГОЭТАПНЫХ КИБЕРАТАК НА ОСНОВЕ АНАЛИЗА ПОВЕДЕНЧЕСКИХ АНОМАЛИЙ С ПРИМЕНЕНИЕМ АЛГОРИТМОВ МАШИННОГО ОБУЧЕНИЯ БЕЗ УЧИТЕЛЯ

Введение. Ускоренное расширение ИТ-инфраструктуры приводит к значительному увеличению массивов телеметрических данных (логи аутентификации, сетевая активность, аудит системного аудита). В таких условиях классические защитные механизмы, базирующиеся на поиск совпадений с заранее заданными сигнатурами, демонстрируют

низкую эффективность при выявлении скрытых многоэтапных атак. Злоумышленники применяют тактику под названием "медленное проникновение" (AdvancedPersistentThreats): получая легитимные учетные записи, они постепенно расширяют свои полномочия, не нарушая формальных политик безопасности. В результате возникает "слепая уязвимость" — формально легитимная активность не вызывает срабатываний сигнатур, что делает подобные угрозы практически невидимыми для стандартных систем защиты. Для закрытия этого пробела предлагается подход, базирующийся на анализе поведенческих аналитики (UEBA) и алгоритмах машинного обучения без учителя. Данный метод позволяет выявлять скрытые угрозы на ранних стадиях без необходимости обновления сигнатурных баз [1,2].

Актуальностьобусловлена экспоненциальным ростом многоэтапных кибератак. Легитимная компрометация учетных записей не вызывает срабатываний сигнатурных системами, создавая слепые уязвимости. Для решения предлагается применение поведенческой аналитики и алгоритмов без учителя. Данный подход выявляет скрытые угрозы на ранних стадиях без обновления сигнатурных баз, однако требует доработки в части снижения ложных срабатываний и интерпретируемости моделей. [3,4].

Подход к обнаружению атак

Методология обнаружения базируется на последовательной трансформации сырых логов в числовые векторы (FeatureEngineering) с последующей их аналитикой. Вычислительный процесс включает четыре этапа:

1. Сбор и очистка данных.Централизованный сбор журналов осуществляется с помощью SIEM-системы Wazuh, развёрнутой в контейнерной среде. Фиксируются события аутентификации (SSH), системного аудита (AppArmor/Kernel) и сетевой активности. Поток данных приводится к единому формату, дубликаты удаляются, события объединяются в логические сессии.[5].

2. Инженерия признаков.Прямой парсинг неструктурированных логов мало пригоден для ML-алгоритмов. Разработан программный модуль на языке Python, который извлекает из сырых данных математические признаки: количество уникальных IP-адресов за сессию, суммарное количество ошибок аутентификации и средний час активности. Это позволяет перевести текстовые события в многомерное пространство, интерпретируемое алгоритмами.

3. Обнаружение аномалий. Для анализа применяется алгоритм IsolationForest (изолирующий лес). Алгоритм строит ансамбль случайных деревьев в многомерном пространстве признаков, формируя «облик» нормального поведения. При поступлении нового вектора алгоритм определяет его изолированность (число разбиений в деревьях решений), что количественно измеряет отклонение от нормы без жёстко заданных порогов.

4. Оценка риска и реакция. На основе степени изолированности рассчитывается индекс риска (RiskScore, от 0 до 100). Высокие значения передаются аналитикам, а для контекстной поддержки принятых решений задействует интегрированная LLM-подсистема, маршрутизация запросов к языковым моделям осуществляется через унифицированный интерфейс OpenRouter API[6,7].

Возникающие проблемы.

Однако не всё так просто. Во-первых, высокий уровень ложных срабатываний — аномалия не всегда означает атаку. Во-вторых, интерпретация результатов моделей остаётся сложной задачей: алгоритм «видит» отклонение, но объяснить его бывает непросто. И, наконец, многое зависит от качества исходных данных.

Плюс есть ещё один фактор — адаптация злоумышленников. Они довольно быстро подстраиваются под новые методы обнаружения. Иногда быстрее, чем хотелось бы.

Заключение. Применение алгоритмов машинного обучения без учителя в связке с инженерией признаков позволяет сместить акцент с реактивным обнаружением угроз на проактивное их предупреждение. Разработанный комплекс доказывает, что открытые SIEM и UEBA-платформы могут быть интегрированы в единую систему, способную выявлять

скрытые аномалии. Дальнейшие исследования должны быть направлены на повышение точности классификации за счёт учета хронологии событий и снижения ложных срабатываний за счёт адаптивных порогов [3,4].

Список литературы

1. KasperskyGlobal. "Отчёт о современных киберугрозах" // Kaspersky. 2026.
2. Positive Technologies. "Ландшафткиберугроз" // Positive Technologies. 2026.
3. SecurityVision. "UEBA: поведенческий анализ пользователей" // SecurityVision. 2024.
4. CloudNetworks. "UEBA-подход к анализу поведения пользователей" // CloudNetworks. 2024.
5. NomanA. "Enhancing IT security with anomaly detection in Wazuh" // Wazuh Blog. 2023.
6. Liu F.T., Zhou Z.-H., Yin J., Huang J.-W., Yu H.-X., Li J.-J., Huang J. "Isolation Forest: Isolation Forest" // arXiv. 2008.
7. OpenRouter. "OpenRouter API Reference" // OpenRouter Docs. 2025.

Столов А.Р., Демьяненко Г.В., Дудкин А.С., Бирюков Д.Н.

Военная-космическая академия им. А.Ф. Можайского, Санкт-Петербург

К ВОПРОСУ О «ПОЛИГРАФИЧЕСКОМ» ТЕСТИРОВАНИИ БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ НА ОСНОВЕ ДИНАМИЧЕСКОГО АНАЛИЗА МЕТРИК ИХ ВНУТРЕННЕГО СОСТОЯНИЯ

Проблема верификации достоверности информации, генерируемой большими языковыми моделями (БЯМ), приобретает критическую значимость в условиях их интеграции в системы поддержки принятия решений, юридические и медицинские сервисы. Традиционные подходы к детекции «лжи» в полиграфических исследованиях базируются на измерении физиологических реакций (частота дыхания, артериальное давление, кожно-гальваническая реакция) при предъявлении вопросов различных категорий [1]. Однако точность таких методов в контролируемых условиях варьируется в пределах 80-90%, а в реальных сценариях существенно снижается из-за влияния индивидуальных психофизиологических особенностей и возможности применения контрмер [2].

Перенос методологических принципов классического полиграфа в область оценки БЯМ открывает новые перспективы: вместо физиологических сенсоров предлагается использовать математические «датчики» внутреннего состояния модели (активации скрытых слоёв, градиенты, распределения внимания), а вместо категорий вопросов – онтологически структурированные сценарии тестирования с динамическим переключением.

Классический полиграф [3] регистрирует изменения вегетативной нервной системы в ответ на стимулы. Ключевой принцип – сравнение реакций на:

- релевантные вопросы (прямые, касающиеся проверяемого события);
- контрольные вопросы (провокационные, вызывающие стресс у невиновного);
- нейтральные вопросы (фоновые, для калибровки).

Динамическое переключение между категориями происходит при обнаружении аномалий в физиологических показателях, что позволяет уточнять гипотезы в реальном времени.

Концепция предлагаемого киберполиграфа заключается в следующем:

– проведение системной аналогии между категориями полиграфических вопросов (релевантные, контрольные, нейтральные) и типами промптов для БЯМ, позволяющей формализовать процесс адаптивного тестирования;

– введение новых показателей, отражающих устойчивость генерации при предъявлении противоречивых, двусмысленных или онтологически некорректных запросов – метрика «когнитивного стресса» модели;

– формирование онтологического каркаса [4] для динамического переключения: автоматическая адаптация тестовых сценариев на основе семантического анализа ответов модели, аналогичная принципу «если модель врет – переключаемся на уточняющую категорию вопросов»;

– применение мультимодальных сенсоров достоверности: разработка архитектуры системы мониторинга, объединяющей внутренние метрики (энтропия активаций, согласованность слоёв) и внешние поведенческие индикаторы (латентность ответа, вариативность генерации при повторных запросах).

Результаты работы предполагают: разработку открытого фреймворка для динамической оценки достоверности БЯМ, снижение частоты галлюцинаций в критических приложениях за счёт раннего детектирования нестабильных генераций, а также формирование методологической основы для стандартизации тестирования ИИ-систем в регулируемых отраслях (медицина, юриспруденция, финансы).

Таким образом, предлагается перенести проверенные принципы полиграфического тестирований в область оценивания больших языковых моделей, заменяя физиологические сенсоры на вычислительные метрики внутреннего состояния, а категории вопросов – на онтологически структурированные сценарии. Основным вкладом видится формализация процесса динамического переключения между типами тестовых запросов к БЯМ на основе многомерного анализа «поведенческих» и «когнитивных» сигналов модели, что открывает путь к созданию адаптивных систем верификации ИИ, способных работать в реальном времени и масштабироваться на различные домены.

Список литературы:

1. Palestra. Value of physiological parameters recorded in the polygraph test procedure // Palestra. 2023. №4. URL: https://palestra.pl/en/palestra/issue/4-2023/article/value-of-physiological-parameters-recordered-in-the-polygraph-test-procedure (дата обращения: 15.01.2026).
2. Morgan Polygraph. The Science of Physiological Responses in Lie Detection. 2024. URL: https://morganpolygraph.com/index.php/2024/12/16/the-science-of-physiological-responses-in-lie-detection-2/ (дата обращения: 15.01.2026).
3. National Research Council. The Polygraph and Lie Detection. Washington, DC: The National Academies Press, 2003. DOI: 10.17226/10420.
4. Garanina N., Anureev I. An Ontology-Based Approach to Support Formal Verification of Concurrent Systems // DataMod. 2019. URL: https://pages.di.unipi.it/datamod/wp-content/uploads/sites/8/2019/10/DataMod_2019_paper_4.pdf

Тельбух В.В.
Военно-космическая академия имени А.Ф. Можайского, Санкт-Петербург

АВТОНОМНАЯ МНОГОСЛОЙНАЯ ПЛАТФОРМА АНАЛИЗА ЦИФРОВОГО СЛЕДА

Введение. Современные киберугрозы характеризуются экспоненциальным ростом открытых данных и сокращением времени появления эксплойтов после публикации CVE (до 5 суток). Одновременно ужесточаются регуляторные требования (152-ФЗ, GDPR), запрещающие передачу конфиденциальной информации внешним облачным LLM. Существующие решения либо собирают данные без семантического анализа (SpiderFoot, Recon-ng и др.), либо требуют GPU-ферм и внешних API (PentestGPT) [4]. Цель работы – создание локальной платформы, которая на потребительском CPU способна выполнять полный цикл: пассивный и агрессивный сбор, поиск утечек, идентификацию CVE, контролируемую эксплуатацию и анонимизацию оператора.

Платформа объединяет 232 модуля, ансамбль из 16 локальных LLM размером ≤9B параметров (12 моделей установлены постоянно, до 4 доступны для оперативной подгрузки), контекстно-зависимый RAG-индекс, каталог из 26 инструментов безопасности с универсальным Python-fallback, а также собственную систему многоагентнойоркестрации на базе LangGraph.

Принципиальными особенностями являются двухрежимная модель эксплуатации (безопасный LabMode и контролируемый BattleMode с криптографической авторизацией каждого критического шага) и шестислойная подсистема анонимизации GhostLayer. Экспериментально подтверждено, что ансамбль малых моделей потребляет на 33% меньше пиковой оперативной памяти (10,5 ГБ против 15,5 ГБ у одиночной 14B-модели Phi-4) и на 64,6% сокращает среднее время агентского шага (18,4 с против 52 с на чистом CPU) при падении качества ответа (ROUGE-L) не более 3%. На корпусе из 200 документов профилированный RAG повысил точность (F1) ответов на вопросы об утечках данных на 24,2%, а в среднем по всем классам – на 19,4%. В модельных экспериментах BattleMode предотвратил несанкционированную эксплуатацию в 100% попыток, а GhostLayer обеспечил прохождение 99,5% запросов через сеть Tor без оставления исходного IP в логах цели и без единого IDS-алерта (Snort). Платформа функционирует на оборудовании уровня IntelCore i7/64 ГБ ОЗУ без GPU, не требует облачных сервисов и пригодна для работы на объектах критической информационной инфраструктуры.

Архитектура. Платформа построена по трёхслойной модели.

Слой сбора данных включает 232 модуля (от безопасного пассивного профиля до полного агрессивного) и универсальный Python-fallback для 26 инструментов безопасности (nuclei, subfinder, sqlmap, gobuster и др.), что позволяет отказаться от WSL/Kali при необходимости – через эмуляцию или обращение к публичным API (NVD, crt.sh, ExploitDB). В тестовой инсталляции слой накопил 104 031 сущность, причём агрессивный профиль дал в 8,4 раза больше данных, чем пассивный.

Слой интеллектуального анализа содержит ансамбль из 16 узкоспециализированных SLM (ruadaptqwen2.5-7B, qwen3-8B, deepseek-r1-8B, codellama-7B, llama-guard3-8B и др.) с адаптивным маршрутизатором, который динамически назначает задачам (обзор документов, анализ уязвимостей, планирование эксплуатации) оптимальные модели по эвристической функции $s(m,t)$, учитывающей приоритеты, доступность и исторические метрики. Контекстно-зависимый RAG-индекс (ChromaDB) размечен по классам данных *data_class*, среди которых выделены утечки данных (leak), техническая информация (technical), корпоративная информация (corporate), персональные данные (personal) и другие. Такая профилизация позволяет агенту автоматически сужать retrieval-контекст и исключать нерелевантные чанки до ранжирования. В слой также интегрированы актуальные базы NVD, MITRE ATT&CK STIX и ExploitDB[1,2]. Для

предотвращения утечек конфиденциальной информации в RAG-индекс и публичные отчёты применяется многоязычный PII-детектор на базе MicrosoftPresidio с 9 кастомными распознавателями (RU/UA/EN), выполняющий анонимизацию на этапах индексации документов и формирования внешнего аудиторского следа [5].

Слой верификации и защиты включает два режима: LabMode, при котором все действия high/critical блокируются (разрешены только безопасные проверки: nuclei --check, sqlmap --level 1 --risk 1, hydra запрещён), и BattleMode, запускаемый только при наличии криптографически подписанного контракта (HMAC-SHA256 + хэш договора). В BattleMode каждый опасный шаг требует поэтапного подтверждения через SSE-интерфейс (тайм-аут 120 с) с возможностью экстренного прерывания emergency-abort (принудительный SIGKILL всех подпроцессов). Параллельно функционирует GhostLayer– шестислойная модель анонимизации, охватывающая сетевой уровень (Tor SOCKS5 с принудительной маршрутизацией), идентификацию (JA3-эмуляция через curl_cffi, пул из 50 000 реальных User-Agent), временной (jitter 0,5–3 с, имитация ночного режима), поведенческий (Referer, DNS-over-HTTPS), хостовой (firejail, изоляция файловой системы) и маскировочный (подмешивание 5% безобидного трафика). Все события записываются в неизменяемый аудиторский след.

Оркестрацию агентов выполняют LangGraphи StateGraph с checkpoints в SQLite, что позволяет приостанавливать и возобновлять многоагентные цепочки без потери промежуточных результатов – это критично при длительных циклах эксплуатации [3]. Многоагентная система включает специализированных агентов: DocumentInvestigatorAgent (поиск секретов, версий ПО, PII, связи с CVE), VulnerabilityAgent, ExploitationAgent (ReAct-цикл с lab_guard), SynthesizerAgent. Коммуникация с внешними инструментами унифицирована через MCP-сервер.

Научная новизна.

1. *Ансамбль малых LLM с адаптивной маршрутизацией.* Впервые показано, что набор специализированных моделей ≤9B заменяет универсальную 14B-модель без значимой потери качества (ΔROUGE-L ≤ 3%), но с радикальным выигрышем в потреблении ресурсов (RAM –33%, время шага –64,6%).

2. *Этически-контролируемая верификация.* Формализован протокол Lab-BattleMode с HMAC-авторизацией и поэтапным одобрением, гарантирующий 100% блокировку несанкционированной эксплуатации. В отличие от аналогов, каждый критический инструмент (sqlmap, nuclei -exploit, metasploit) в BattleMode ожидает живого подтверждения.

3. *Интегрированный GhostLayer.* Шестислойная модель анонимизации, реализованная как единая подсистема, обеспечивает скрытность оператора (99,5% трафика через Tor, 0 IDS-алертов) и защиту от подставных заказов без потери полного аудиторского следа внутри платформы.

4. *Метод контекстно-зависимого RAG с профилированием по классам данных.* Фильтрация на уровне класса данных (leak, technical, corporate, personal) повышает F1-меру ответов на вопросы по утечкам на 24,2% при сокращении времени retrieval на 57,1%. Наибольший прирост дают профильные запросы (leak), где без профилирования модель часто смешивает корпоративные и персональные данные.

5. Универсальный Python-fallback для 26 инструментов. Впервые все ключевые утилиты безопасности имеют встроенную Python-реализацию (через публичные API или эмуляцию), что делает платформу работоспособной даже без специализированных бинарных сборок и WSL.

Экспериментальная оценка. Эксперименты проводились на стенде IntelCore i7-13700H, 64 ГБ ОЗУ, без GPU, под управлением Windows 11 и WSL2 для инструментов, требующих Linux-окружения.

Сравнение ансамбля SLM и универсальной модели Phi-4 (14B) на 50 идентичных агентских шагах: ансамбль показал пиковое потребление ОЗУ 10,5 ГБ (против 15,5 ГБ у

14B-модели), среднее время шага 18,4 с (против 52 с), долю успешных вызовов инструментов 94% (против 90%) при ROUGE-L 0,71 (против 0,73). Снижение времени достигнуто за счёт меньшего объёма параметров и узкой специализации каждой модели, сокращающей число итераций ReAct-цикла.

Работа RAG с профилированием: на корпусе из 200 документов F1 выросла с 0,617 до 0,737. Наибольший прирост достигнут на leak-вопросах (+24,2%), на технических данных +19,0%, на корпоративных +15,4%, что объясняется устранением нерелевантных чанков до ранжирования.

Тестирование BattleMode: в 150 попытках запуска sqlmaplevel=3, nuclei -exploit и hydra без авторизации все были заблокированы (100%); в 30 легитимных Battle-сессиях отказов не зафиксировано.

Испытания GhostLayer: из 2000 запросов 1990 прошли через Tor (99,5%), ни один исходный IP не раскрыт, сигнатуры Snort не выдали ни одного характерного для сканирования алерта. Оверхед по времени составил 42,7% – допустимый результат за полную анонимизацию.

Атрибуция угроз: автоматический маппинг 47 CVE на техники MITRE ATT&CK моделью mistralai-7B достиг точности 89,5%.

Полное регрессионное тестирование (191 тест, 0 отказов) подтверждает стабильность всех заявленных функций.

Заключение. Разработанная архитектура доказывает, что объединение ансамбля малых LLM с жёстким этическим контролем и сквозной анонимизацией позволяет создать полнофункциональную Red/BlueTeam платформу, пригодную для работы в изолированных сетях и на объектах КИИ. Платформа не требует облачных подключений или GPU, полностью соответствует отечественным нормативным требованиям и может быть развёрнута на стандартном рабочем месте аналитика. Дальнейшие исследования направлены на внедрение механизмов долговременной памяти агентов и динамической генерации compositeattackchains с использованием reasoning-моделей.

Список литературы:

1. MITRE ATT&CK v16. Enterprise Matrix, 2024.
2. MITRE STIX 2.1 Specification, OASIS, 2021.
3. LangGraphDocumentation, LangChain, 2024.
4. SpiderFoot 4.0 Documentation, 2024.
5. MicrosoftPresidioAnalyzer, 2024.

Тельбух В.В.

Военно-космическая академия имени А.Ф. Можайского, Санкт-Петербург

ИНТЕЛЛЕКТУАЛЬНАЯ ПЛАТФОРМА ДЛЯ АНАЛИЗА И ЭКСПЕРТИЗЫ УЧЕБНО-МЕТОДИЧЕСКОЙ ДОКУМЕНТАЦИИ НА ОСНОВЕ ОРКЕСТРАЦИИ АНСАМБЛЯ МАЛЫХ LLM И ДВУХКОНТУРНОГО RAG-КОНТРОЛЯ КАЧЕСТВА

Современный вуз сопровождает десятки тысяч учебно-методических документов, качество которых напрямую определяет результаты обучения. Полная ручная экспертиза одной рабочей программы дисциплины (РПД) силами методиста занимает до 8 часов, а фронтальный аудит всего фонда практически неосуществим. При этом существующие ИИ-решения либо решают узкую задачу (заимствования, прокторинг), либо остаются на уровне концептов. Цель работы – создание архитектурно организованной интеллектуальной среды, способной автоматизировать анализ, структурирование, контроль качества и экспертную оценку всего спектра учебно-методической документации

образовательной организации за счёт технологий Retrieval-AugmentedGeneration (RAG – генерация с дополнением поиском) [1] и оркестрации ансамбля компактных языковых моделей.

Постановка задачи и сравнение с аналогами. Платформа должна охватывать все основные типы документов: квалификационные требования, учебные планы, матрицы компетенций, рабочие программы дисциплин с их разделами (тематический план, тексты лекций, фонды оценочных средств, методические указания, списки литературы и ПО), программы практик и междисциплинарные экзамены.

Архитектурное решение. Платформа построена на стеке Python, Flask, Ollama, ChromaDB и реализует девятиуровневую модульную архитектуру. Ядро – восьмистадийный RAG-конвейер. Запрос нормализуется, классифицируется (TaskClassifier, точность 94,2%, F1=0,94) и маршрутизируется к одной из специализированных языковых моделей (LLM). Семантический поиск выполняется в векторной базе ChromaDB [4,5] с фильтрацией по типу документа. Интеллектуальный роутер выбирает эффективнуюмодельдля работы анализатора: codellama:7b (валидация часов), llama3.1:8b (структурный анализ, дубликаты), qwen2.5:14b (ФГОС, компетенции). Все модели – 4-битное квантование на основе архитектуры LLaMA [3]. Благодаря политике «одна активная модель одновременно» пиковое потребление ОЗУ не превышает 14 ГБ (достаточно сервера 64 ГБ без графического ускорителя).

Одиннадцать анализаторов закрывают все типы документов. Каждый из них проверяет ключевые элементы учебного процесса – от содержания и структуры до нагрузки и компетенций:

- **анализатор соответствия ФГОС** (FGOSAnalyzer) – проверяет, охватывает ли содержание дисциплины все обязательные компетенции согласно федеральным стандартам;
- **анализатор компетенций** (CompetenciesAnalyzer) – сопоставляет матрицу компетенций с рабочими программами, выявляя избыточность или пробелы;
- **анализатор структуры** (StructureAnalyzer) – контролирует наличие всех обязательных разделов РПД: от пояснительной записки до фонда оценочных средств;
- **анализатор трудоёмкости и часов** (WorkloadAnalyzer) – валидирует распределение часов между видами занятий и общую трудоёмкость;
- **анализатор дублирования** (DuplicatesAnalyzer) – выявляет повторы тем и материалов как внутри одной дисциплины, так и между смежными курсами;
- **анализатор междисциплинарных связей** (SequenceAnalyzer) – определяет логическую преемственность дисциплин по компетенциям и содержанию;
- **анализатор содержания** (ContentAnalyzer) – оценивает дидактическую глубину текстов лекций и иных учебных материалов;
- **анализатор литературы** (LiteratureAnalyzer) – проверяет актуальность и релевантность рекомендованных источников;
- **анализатор видов учебных занятий** (SessionTypeAnalyzer) – комплексно анализирует лекции, семинары, практические и лабораторные работы, курсовое проектирование и самостоятельную работу;
- **анализатор общего качества** (QualityAnalyzer) – сводит результаты всех проверок в единую интегральную оценку;
- **анализатор произвольных запросов** (GeneralAnalyzer) – обеспечивает свободный семантический поиск по всему корпусу документов.

Каждый анализатор использует специализированный промпт с раздельными параметрами генерации, что гарантирует точность и релевантность заключений.

Двухконтурная проверка качества.Уникальная особенность архитектуры – двухконтурный контроль качества (DocumentQualityValidator). Идея метаоценки ответа опирается на подход RARR (RetrofitAttributionusingResearchandRevision) [2], однако в

платформе она реализована в двух раздельных контурах. *Первый конт ур* – валидация исходного учебного документа по пяти критериям (парсинг, полнота, очистка, структура, согласованность). *Вт орой конт ур* – метаоценка аналитического заключения: проверка ссылок, непротиворечивости и формата. Именно этот контур свёл долю критических ошибок в ответах с 5,3% до **0%** на 150 тестовых запросах.

Значение для участников образовательного процесса. Внедрение платформы напрямую повышает качество обучения. Для **обучающихся** гарантируется актуальность и непротиворечивость осваиваемого материала, исчезают дублирования и пробелы, оценочные средства точно соответствуют заявленным компетенциям. **Преподаватель** получает оперативную обратную связь по своим РПД и текстам лекций, видит конкретные рекомендации и может самостоятельно поддерживать документацию в эталонном состоянии, тратя на рутину в 7,5 раз меньше времени. **Учебно-методический отдел** обретает инструмент фронтального аудита всей документации направлений, формирует консолидированные отчёты для учебно-методических советов и снижает нагрузку на методистов. **Заведующие кафедрами** получают прозрачную аналитическую картину состояния учебно-методического фонда, могут объективно оценивать его качество и принимать обоснованные кадровые решения. **Руководство ВУЗА** использует агрегированные метрики как для внутреннего управления, так и для внешней отчётности, в том числе при аккредитации. Ролевая модель управления доступом (RBAC – Role-BasedAccessControl) гарантирует, что каждый участник видит именно те документы и функции, которые соответствуют его должностным обязанностям. Все данные обрабатываются локально, что снимает риски утечки служебной информации.

Экспортный слой и готовые отчёты. Результаты анализа не просто отображаются на экране – они агрегируются модулем построения отчётов (AnalysisReportBuilder) и формируются как готовые документы в форматах PDF, DOCX и JSON. Каждый отчёт содержит разделы «Общая оценка», «Выявленные проблемы», «Детализация» и «Рекомендации», что позволяет выносить его на учебно-методический совет сразу после генерации, без дополнительной ручной обработки. Этот слой замыкает цикл «анализ → решение», делая платформу полноценным инструментом управления качеством.

Экспериментальная верификация. Эксперимент проведён на уровне кафедры. Проанализированы 14 дисциплин (более 1400 документов). Стенд – сервер с 64 ГБ ОЗУ без графического ускорителя (CPU-only). Сравнивались ансамбль с роутером и универсальная модель qwen2.5:32b (всес 4-битным квантованием, одинаковые промпты). Качество оценивали три методиста (αКронбаха 0,87).

Таблица 1. Средняя экспертная оценка качества ответов (5-балльная шкала)

Тип анализа	llama3.1:8b	qwen2.5:32b	Ансамбль + роутер	Прирост ансамбля к 32B, %
Валидация часов	2,8±0,3	3,9±0,2	4,7±0,2	+20,5
Структурный анализ	3,5±0,3	4,2±0,2	4,4±0,2	+4,8
Компетенции	3,3±0,3	4,3±0,2	4,5±0,2	+4,7
Дублирование	3,9±0,2	4,1±0,2	4,6±0,2	+12,2
Анализ литературы	4,0±0,2	4,3±0,2	4,8±0,2	+11,6
Среднее	**3,50±0,3**	**4,16±0,2**	**4,60±0,2**	**+10,6**

Различия статистически значимы (критерий Уилкоксона, $p < 0,01$; 95% доверительный интервал разницы средних: +0,34…+0,54 балла).

Таблица 2. Производительность и надёжность (CPU)

Показатель	llama3.1:8b	qwen2.5:32b	Ансамбль + роутер
Среднее время ответа (один запрос), с	4,8	48,0	6,0
Пиковое потребление ОЗУ, ГБ	10	24	14
Среднее время ответа при 10 польз., с	7,2	н/д*	9,8
Критические ошибки до валидатора	12,0% (18/150)	6,0% (9/150)	5,3% (8/150)
Критические ошибки после валидатора	1,3% (2/150)	0,7% (1/150)	**0% (0/150)**

* Нагрузочное тестирование для модели 32B не проводилось.

Научная новизна и практическая значимость. Впервые предложен и верифицирован принцип оркестрации малых специализированных языковых моделей (LLM) с доказанным превосходством над универсальной моделью размера 32B. Ключевые архитектурные компоненты – классификатор запросов (точность 94,2%), интеллектуальный роутер, 11 анализаторов и двухконтурный контроль качества – формируют замкнутый цикл от загрузки документа до аудируемого отчёта, сокращая время экспертизы в 7,5 раз и полностью устраняя критические ошибки. Полная локальная обработка данных и работа на CPU-only сервере делают платформу доступной любой кафедре уже сегодня, а в перспективе учебному заведению в целом.

Интеграция в образовательный процесс. Эксперимент на уровне кафедры подтвердил жизнеспособность подхода и выявил точки масштабирования. Архитектура платформы изначально проектировалась как встраиваемая в информационную среду ВУЗА. Поддержка инкрементальной переиндексации позволяет синхронизироваться с ежегодным обновлением РПД, а открытый REST API – интегрироваться с действующими системами управления обучением и прочими системами документооборота. По мере подключения новых кафедр и накопления данных платформа формирует цифровой профиль качества каждой дисциплины и образовательной программы, что даёт руководству объективную основу для принятия решений о корректировке содержания, перераспределении нагрузки и развитии кадрового состава. Именно такой подход – от пилотной кафедры к полномасштабному внедрению – превращает разовую экспертизу в непрерывный процесс управления качеством образования, прямо влияющий на уровень подготовки выпускников.

Список литературы

1. Lewis P., Perez E., Piktus A., Petroni F. Et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks // Advances in Neural Information Processing Systems 33 (NeurIPS 2020). – 2020. – Pp. 9459–9474.
2. Gao L., Dai Z., Pasupat P., Chen A. et al. RARR: Researching and Revising What Language Models Say, Using Language Models // Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL 2023). – 2023. – Pp. 16477–16508.
3. Touvron H., Lavril T., Izacard G., Martinet X. et al. LLaMA: Open and Efficient Foundation Language Models // arXiv preprint arXiv:2302.13971. – 2023.

4. Reimers N., Gurevych I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks // Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP-IJCNLP 2019). – 2019. – Pp. 3980–3990.
5. Chroma: The Open-source AI Application Database. – URL: https://docs.trychroma.com/ (датаобращения: 07.05.2026).

Толгоренко Е.М., Зубков Е.А., Овасапян Т.Д.

C

ОЦЕНКА ПРИМЕНИМОСТИ БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ ДЛЯ ПРОТИВОДЕЙСТВИЯ АТАКЕ ПОДДЕЛКИ ЛИЧНОСТИ

Атаки подделки личности в корпоративной среде относятся к числу наиболее опасных форм социальной инженерии, поскольку злоумышленник действует от лица доверенного сотрудника и опирается не на технические уязвимости, а на контекст общения и доверие адресата. В таких сценариях традиционные средства защиты, ориентированные на вложения, ссылки, домены и прочие технические признаки, часто оказываются недостаточно эффективными, особенно при коммуникации через мессенджеры.

Для выявления факта подделки личности предлагается анализ текстов сообщений, поскольку данный способ применим для широкого круга сценариев атаки (от использования поддельных аккаунтов до использования скомпрометированных аккаунтов). Конкретно для выявления атаки подделки личности предполагается анализ стиля письма автора сообщения. На основе анализа делается вывод о том, был ли текст действительно написан человеком, за которого себя выдает автор сообщения (задача подтверждения авторства).

Методы, применяемые для подтверждения авторства в контексте анализа стиля письма, используют два типа признаков – стилометрические признаки или сам текст.

Стилометрические признаки представляют собой количественные характеристики стиля письма. Признаки включают несколько групп параметров, при этом вычисление некоторых из них не представляет сложности (частота знаков пунктуации, используемые эмодзи, количество ссылок), тогда как вычисление других является трудоемким (упоминаемые именованные сущности, использование диалектизмов или архаизмов, грамматические ошибки).

Определять принадлежность текста автору можно также путем анализа самого текста с помощью большой языковой модели, обученной на текстах этого автора. Для этого используется перплексия – мера ошибки предсказания следующего токена большой языковой моделью. В базовом виде перплексия текста представляет собой усредненную ошибку для всех токенов текста, что снижает точность результатов модели. Для повышения точности результатов на основе перплексии каждого отдельного токена строится вектор, включающий такие признаки, как:

– базовое значение перплексии токена;
– энтропия распределения перплексии;
– максимальное значение перплексии текста;
– ранг следующего токена (положение следующего токена в списке, упорядоченном по убыванию вероятностей токенов);
– суммарная вероятность токенов с большей или равной вероятностью, чем у следующего токена.

Для учета этих признаков для больших языковых моделей добавляется выходной слой, обучающийся на тренировочной выборке. Схема работы метода представлена на рисунке 1.

u
m
e
m

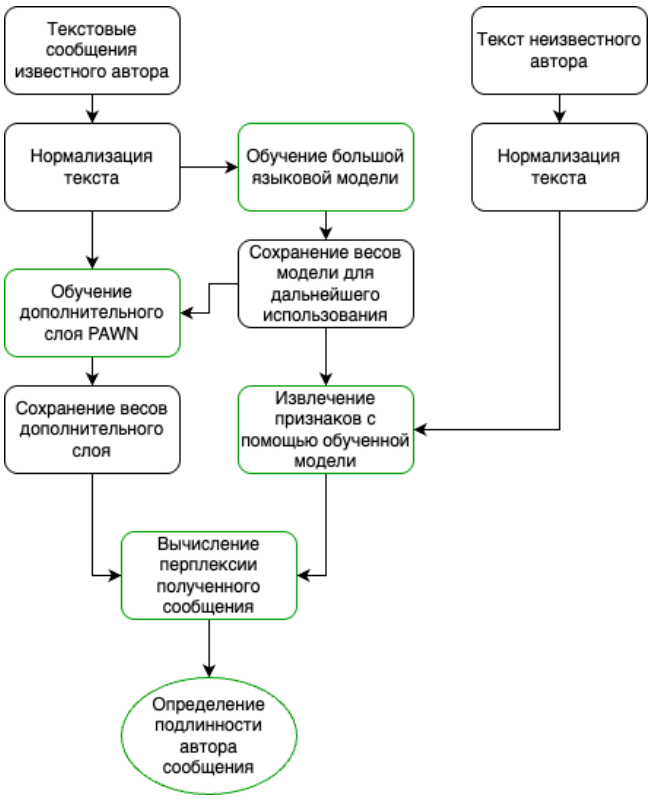


Рисунок 1 – Схема работы предлагаемого метода

Точность данного метода была проверена на наборе данных, составленном из 55 034 личных сообщений в мессенджере от 10 различных авторов. На основе текстов для каждого из авторов были обучены по 2 большие языковые модели, за основу были взяты модели семейства Qwen2.5 размера 0,5 млрд и 1,5 млрд параметров. Результаты обучения моделей представлены в таблице 1.

Таблица 1 – Результаты тестирования моделей

Автор	Количество текстов для тестирования	Precision		Recall		F1		MCC		BA	
		0,5B	1,5B	0,5B	1,5B	0,5B	1,5B	0,5B	1,5B	0,5B	1,5B
user-1	2 621	**0,65**	**0,65**	**0,91**	**0,88**	**0,75**	**0,75**	**0,44**	**0,43**	**0,70**	**0,70**
user-2	831	0,35	0,41	0,56	0,51	0,43	0,45	0,31	0,35	0,69	0,69
user-3	669	0,47	0,48	0,46	0,41	0,46	0,44	0,39	0,38	0,70	0,68
user-4	498	0,42	0,42	0,49	0,41	0,45	0,41	0,40	0,37	0,72	0,68
user-5	259	0,59	0,35	0,28	0,27	0,38	0,30	0,39	0,28	0,64	0,63
user-6	236	0,40	0,27	0,22	0,36	0,28	0,31	0,27	0,27	0,60	0,66
user-7	121	0,22	0,17	0,27	0,32	0,24	0,22	0,23	0,22	0,63	0,65
user-8	104	0,46	0,58	0,31	0,27	0,37	0,37	0,36	0,39	0,65	0,63
user-9	89	0,36	0,24	0,09	0,08	0,14	0,12	0,18	0,13	0,54	0,54
user-10	75	0,60	0,58	0,18	0,33	0,28	0,42	0,32	0,43	0,59	0,67

При использовании полученных моделей для подтверждения авторства удалось добиться точности, равной 0,75. При этом чем меньше текстов автора было доступно для обучения, тем меньше получалась точность модели. Также было определено, что при анализе коротких сообщений размер модели незначительно влияет на точность определения.

Черникова В.М., Калинин М.О.
Санкт-Петербургский политехнический университет Петра Великого, Санкт-Петербург

ОБНАРУЖЕНИЕ СОСТЯЗАТЕЛЬНЫХ АТАК НА СИСТЕМЫ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА НА ОСНОВЕ ВАРИАЦИОННОГО АВТОЭНКОДЕРА И ОЦЕНКИ ПРАВДОПОДОБИЯ ВХОДНЫХ ДАННЫХ

Одной из ключевых угроз безопасности систем искусственного интеллекта (ИИ) являются состязательные атаки, включающие искажения и уклонения вычислительных моделей принятия решений, при которых во входные данные вносятся малые возмущения, способные существенно снизить точность и достоверность генерируемых прогнозов [1, 2].

Существующие методы защиты от состязательных атак можно разделить на несколько видов: методы, модифицирующие процесс обучения модели (состязательное обучение, защитная дистилляция), методы предобработки входных данных (аугментация, сжатие признаков), а также методы, основанные на выявлении отклонений входных данных от обучающей выборки. К последним относится метод MagNet [3], в котором реализуются механизмы детектирования состязательных примеров на основе оценки ошибки реконструкции с помощью автоэнкодеров. Большинство известных решений обладают рядом ограничений. Состязательное обучение требует значительных вычислительных ресурсов и неустойчиво при адаптивных атаках, учитывающих свойства защиты. Методы предобработки и сжатия признаков эффективны только против слабых возмущений. Инструментарий MagNet, несмотря на свою гибкость, использует детерминированные автоэнкодеры, что снижает устойчивость к адаптивным атакам, таких как DeepFool и C&W. Кроме того, большинство методов защиты снижают точность прогнозов на корректных входных данных.

На базе метода MagNet авторами разработан метод обнаружения состязательных атак, использующий вариационный автоэнкодер и оценивающий не только ошибку реконструкции, но и правдоподобие входных примеров в выборке на основе скрытого вероятностного представления данных. Такой подход позволяет анализировать соответствие входных примеров обученному распределению. В методе MagNet выявление искаженных данных осуществляется на основе следующих метрик: ошибки реконструкции (характеризует отличие восстановленного изображения от исходного) и дивергенции Йенсена-Шеннона JSD (измеряет отклонение от априорного распределения). В методе MagNet метрики сравниваются линейно с пороговыми значениями. В разработанном решении список признаков расширен и ранжирован с помощью метода «случайного леса» (Random Forest), при этом, нелинейно. Внедрены следующие метрики:

1. Ошибки реконструкции: $rec_p = \|x - vae(x)\|_p$, где x – входные данные, $vae(x)$ – реконструированный пример, p – вид нормы $p = 1$ и $p = 2$.
2. Параметры латентного пространства: среднее значение μ, отклонение σ^2;
3. Нижняя граница правдоподобия $ELBO = rec_p + D_{KL} = rec_p + \left(-\frac{1}{2}(log\sigma^2 - \sigma^2 - \mu^2 + 1)\right)$;
4. Дивергенция Йенсена-Шеннона $JSD\ (P||Q) = \frac{1}{2}D_{KL}(\mathrm{P} \parallel \mathrm{M}) + \frac{1}{2}D_{KL}(\mathrm{Q} \parallel \mathrm{M})$, где $D_{KL}(\mathrm{P} \parallel Q) = \sum_i P(i)\ log\ \frac{P(i)}{Q(i)}$, $M = \frac{1}{2}(P + Q)$.

Обнаружение состязательных атак на системы ИИ на основе вариационного автоэнкодера и оценки правдоподобия входных данных выполняется следующим образом:

1. Пример подается на вход вариационного автоэнкодера и вычисляются метрики.
2. Данные передаются в детектор Random Forest и принимается решение о наличии искажения.
3. К состязательным примерам применяется реконструкция со степенью уверенности детектора.

Реализован экспериментальный стенд и проведено сравнение предложенного метода с архитектурой MagNet на наборе данных MNIST. Полученные результаты показали, что решение позволяет повысить способность выявления состязательных атак за счет анализа латентного распределения данных, при этом сохраняется качество классификации на корректных входных примерах (табл. 1).

Таблица 1. Сравнение метода MagNet и разработанного метода

Тип атаки	ε	Метрика	Целевой классификатор	MagNet	Разработанный метод
Исходные данные	–	*Accuracy*	0,991	0,99	0,993
		F1	0,991	0,99	0,993
pgd_linf	0,005	*Accuracy*	0,99	0,99	0,99
		F1	0,99	0,99	0,99
pgd_linf	0,01	*Accuracy*	0,98	0,98	0,98
		F1	0,98	0,98	0,98
pgd_l2	0,5	*Accuracy*	0,957	0,957	0,964
		F1	0,956	0,956	0,963
pgd_l2	1	*Accuracy*	0,908	0,914	0,93
		F1	0,908	0,914	0,93
Deepfool	–	*Accuracy*	0,663	0,663	0,701
		F1	0,664	0,662	0,70

Разработанный метод превосходит передовые известные решения за счет отсутствия необходимости переобучения целевой системы ИИ, меньших вычислительных затрат, а также полного сохранения качества классификации на корректных входных примерах, в отличие от подходов, приводящих к ухудшению качества на чистых данных.

Список литературы:

1. Felix Viktor Jedrzejewski, Lukas Thode, Jannik Fischbach, Tony Gorschek, Daniel Mendez, Niklas Lavesson, Adversarial Machine Learning in Industry: A Systematic Literature Review / Computers & Security. – 2024. – №145.
2. Кириллов Р.Б., Калинин М.О. Выявление искажающих данных в системах обнаружения вторжений, использующих вычислительные модели машинного обучения / Проблемы информационной безопасности. Компьютерные системы. – 2025 – № 1. – С. 59–68.
3. Dongyu Meng, Hao Chen. MagNet: a Two-Pronged Defense against Adversarial Examples / arXiv:1705.09064, 2017.

Шелухин О.И., Маторин Ф.А.

Московский технический университет связи и информатики (МТУСИ), Москва

АУГМЕНТАЦИЯ СЕТЕВОГО ТРАФИКА НА ОСНОВЕ ДИФФУЗИОННОЙ МОДЕЛИ TABDDPM В ЗАДАЧАХ ОБНАРУЖЕНИЯ ВТОРЖЕНИЙ

Публикация выполнена в рамках гранта на реализацию отраслевой научно-педагогической школы МТУСИ "Современные технологии исследования аномалий в информационной безопасности" по проекту "Обнаружение и прогнозирование редких аномальных событий для обеспечения информационной безопасности" (Пр. 93-х от 25.04.2025).

Недостаток размеченного сетевого трафика остаётся одной из причин снижения качества систем обнаружения вторжений: редкие атаки и сценарии «нулевого дня»

представлены слабо, а публикация реальных трасс ограничена требованиями конфиденциальности. Поэтому аугментация данных рассматривается не как простое копирование наблюдений, а как управляемое расширение обучающей выборки за счёт синтетических примеров, сохраняющих статистику реального трафика. Для практической реализации аугментации используется табличная диффузионная модель TabDDPM, позволяющая порождать новые записи сетевых сессий и затем проверять их реалистичность по согласованному набору критериев [1, 2].

В предлагаемой постановке аугментация понимается как вероятностное порождение новых сетевых сессий, сохраняющих форму распределений, хвостовые области и межпризнаковые связи исходного трафика. Для этого используется модель TabDDPM, в которой числовые признаки описываются гауссовой диффузией, а категориальные и бинарные признаки в общей постановке могут описываться полиномиальной диффузией. Важно, что генератор не копирует наблюдения, а обучается обратному переходу от шума к данным.

Диффузионная модель задаётся прямым и обратным процессами. Обратный процесс является марковской цепью с обучаемыми переходами, начинающейся с заданного шумового распределения: $p_\theta(x_{0:T}) = p(x_T)\prod_{t=1}^{T} p_\theta\,(x_{t-1} \mid x_t), \quad p(x_T) = N(x_T; 0, I)$, где: $x_{0:T}$ — траектория от исходного объекта x_0 до шумового состояния x_T; $p_\theta(x_{t-1} \mid x_t)$ — обучаемый обратный переход; $N(x_T; 0, I)$ — стандартное нормальное распределение, из которого начинается генерация.

Прямой процесс при обучении фиксирован. Он последовательно добавляет гауссов шум к объекту x_0 и переводит его в последовательность состояний $x_{1:T}$. В данной постановке этот процесс задаётся не одной операцией, а произведением одношаговых переходов:

$$q(x_{1:T} \mid x_0) = \prod_{t=1}^{T} q\,(x_t \mid x_{t-1}), \quad q(x_t \mid x_{t-1}) = N\big(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I\big),$$

где: $q(x_{1:T} \mid x_0)$ — прямой процесс зашумления; $q(x_t \mid x_{t-1})$ — одношаговый переход; β_t — дисперсия добавляемого шума; I — единичная матрица.

Затем модель обучается обратному гауссову переходу:

$$p_\theta(x_{t-1} \mid x_t) = N\big(x_{t-1}; \mu_\theta(x_t, t), \Sigma_\theta(x_t, t)\big),$$

где: $\mu_\theta(x_t, t)$и $\Sigma_\theta(x_t, t)$ — математическое ожидание и ковариация обратного перехода, определяющие получение менее зашумлённого состояния x_{t-1} из x_t.

Ключевой алгоритмический смысл DDPM состоит в том, что среднее обратного перехода выражается через предсказание шумовой компоненты [3]. Поэтому для числовой ветви TabDDPM скоринговая сеть фактически обучается восстанавливать шум ε, добавленный в прямом процессе:

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\varepsilon_\theta(x_t, t)\right);\ L_t^{simple} = \mathbb{E}_{x_0,\varepsilon,t}\big[\|\varepsilon - \varepsilon_\theta(x_t, t)\|_2^2\big].$$

Здесь α_t=1−β_t, $\bar{\alpha}_t$ — накопленное произведение α_i при i ≤ t; $\varepsilon_\theta(x_t, t)$ — предсказанный шум, ε — истинный шум, а L_t^{simple} задаёт ошибку их согласования. Точное предсказание ε позволяет восстановить реалистичную запись сетевой сессии.

Исследования показали, что для оценки качества аугментации целесообразно использовать набор критериев. Так метрика Колмогорова-Смирнова KS и расстояние Вассерштейна WD позволяют оценивают близость одномерных распределений по каждому признаку и затем усредняются по d признакам [2]:

$$D_j = \sup_x\big|F_j(x) - \hat{F}_j(x)\big|; \quad \text{KS} = \frac{1}{d}\sum_{j=1}^{d} D_j\,;\ \text{WD} = \frac{1}{d}\sum_{j=1}^{d}\int_0^1\big|F_j^{-1}(u) - \hat{F}_j^{-1}(u)\big|\,du,$$

где: $F_j(x)$ и $\hat{F}_j(x)$ — эмпирические функции распределения реальных и синтезированных данных по j-му признаку; D_j — их максимальное расхождение; d — число признаков, а $F_j^{-1}(u)$ и $\hat{F}_j^{-1}(u)$ — соответствующие обратные функции распределения.

Для проверки сохранения межпризнаковых связей целесообразно использовать среднее абсолютное расхождение парных корреляций, а охват опорной области реальных

данных проверяется попаданием синтетических значений в диапазон между 1-м и 99-м процентилями:

$$\Delta\text{Corr} = \frac{2}{d(d-1)} \sum_{1 \le j < k \le d} \left| \rho_{jk}^{(\text{real})} - \rho_{jk}^{(\text{synth})} \right|.$$

$$cov_j(W_k) = \frac{1}{|W_k|} \sum_{x \in W_k} \mathbf{1}\, F_j^{-1}(0.01) \le x^{(j)} \le F_j^{-1}(0.99).$$

$$\text{Coverage}(W_k) = \frac{1}{d} \sum_{j=1}^{d} c\, ov_j(W_k).$$

$$\text{AUC} = \int_0^1 \text{TPR}\left(\text{FPR}^{-1}(t)\right) dt.$$

где: TPR — доля правильно распознанных синтетических объектов; FPR — доля реальных объектов, ошибочно отнесённых к синтетическим; интеграл задаёт площадь под ROC-кривой.

Для объединения разнонаправленных метрик можно использовать модифицированные частные критерии: $K_1 = 1 - 2|\text{AUC} - 0.5|$, $K_2 = 1 - \text{KS}$, $K_3 = 1 - \text{WD}$, $K_4 = 1 - \Delta\text{Corr}$, $K_5 = \text{Coverage}$ и их сумму $\sum_{i=1}^{5} K_i = K_1 + K_2 + K_3 + K_4 + K_5$.

Здесь K_1 — индекс неотличимости: он равен 1 при AUC=0,5 и уменьшается при росте различимости. K_2, K_3 и K_4 переводят ошибки распределений и корреляций в шкалу «чем больше, тем лучше», K_5 сохраняет смысл Coverage, а $\sum K_i$ объединяет пять критериев.

На примере трафика android-приложений [4] показано, что модель TabDDPM позволяет формировать наборы синтетических данных, статистически близкие к реальным данным по одномерным распределениям и совокупностям признаков. Одномерные распределения ключевых атрибутов демонстрируют устойчивое совпадение по положению мод и основной массе. t-SNE-проекции подтверждают перекрытие кластеров реальных и сгенерированных моделью точек, отсутствие существенных систематических смещений центральных квантилей и сохранение хвостов распределений, в том числе в областях редких режимов работы приложений. Показано, что устойчивое качество генерации достигается уже при объёме порядка 1,5–2,5 тыс. сессий на приложение, чего достаточно для построения обучающих выборок, стресс-сценариев и настройки порогов детекторов вторжений в условиях дефицита размеченного трафика на практике. Это позволяет применять синтез выбранных приложений с помощью модели TabDDPM для дальнейшей настройки и валидации детекторов, ориентированных на появление неизвестных классов.

Список литературы:

1. Yang L., et al. Diffusion models: A comprehensive survey of methods and applications // ACM Computing Surveys. 2024. Vol. 56. Iss. 4. PP. 105:1–105:39. DOI:10.1145/3626235
2. Шелухин О. И., Маторин Ф. А. Генерация реалистичного синтезированного сетевого трафика android-приложений с помощью диффузионной модели TABDDPM // В
3. Kotelnikov A., Baranchuk D., Rubachev I., Babenko A. TabDDPM: Modelling tabular data with diffusion models // arXiv preprint arXiv:2209.15421. 2022. DOI:10.48550/arXiv.2209.15421
4. Шелухин О. И., Маторин Ф. А. Снижение размерности массивов данных с помощью многослойных автокодировщиков в задаче классификации мобильных приложений // Труды учебных заведений связи. – 2024. – Т. 10, № 6. – С. 111-120.

Шелухин О.И., Осин А.В., Хализев К.А.
Московский технический университет связи и информатики, Москва

ОБНАРУЖЕНИЕ РЕДКИХ АНОМАЛЬНЫХ СОБЫТИЙ, СВЯЗАННЫХ С ПОВЕДЕНИЕМ ИИ-АГЕНТОВ ПРИ ВЗАИМОДЕЙСТВИИ С ИНФОРМАЦИОННЫМИ СИСТЕМАМИ

Развитие ИИ-агентов на основе больших языковых моделей приводит к изменению классической постановки задачи обнаружения аномального поведения в информационных системах. Если ранее основным объектом анализа являлся пользователь, выполняющий действия через интерфейс системы, то теперь часть операций может быть передана программному агенту. Такой агент способен получать высокоуровневую задачу, самостоятельно декомпозировать её на шаги, выбирать инструменты и взаимодействовать с внешними сервисами, базами данных и API [1, 2]. В результате субъектом взаимодействия с информационной системой становится не только человек, но и алгоритмическая система, поведение которой зависит от модели, управляющего запроса, контекста, доступных инструментов и накопленного состояния.

Агентская модель повышает скорость выполнения рутинных операций, однако одновременно создаёт новые риски информационной безопасности. Если обычный пользователь действует сравнительно медленно и в пределах привычного интерфейса, то ИИ-агент может выполнять цепочки действий автоматически и масштабно. При наличии расширенных полномочий он способен обратиться к данным, не относящимся к исходной задаче, вызвать опасный инструмент, сформировать некорректный запрос к базе данных, изменить состояние системы или передать чувствительные сведения во внешний сервис. Поэтому контроль только итогового ответа агента недостаточен: потенциально опасные события могут возникать непосредственно в траектории достижения результата. Подобные риски автономных LLM-агентов и их траекторий выполнения задач рассматриваются в современных исследованиях [3, 4].

Под редкими аномальными событиями, связанными с поведением ИИ-агентов, предлагается понимать не любые необычные действия, а такие отклонения от ожидаемой траектории выполнения задачи, которые потенциально способны причинить вред информационной системе, данным, пользователям или бизнес-процессам.

Например, аномальным событием может быть обращение агента к административному API при задаче информационного поиска, попытка записи в базу данных при сценарии «только чтение», передача персональных данных во внешний сервис, нарушение установленной последовательности рабочего процесса или переход к неразрешённому источнику данных. Подобные события могут встречаться редко, однако обладать высоким потенциальным ущербом, поэтому должны рассматриваться как приоритетный объект обнаружения.

П

у

с

т

ь

и

м

е

е

т

с

я

Для обнаружения редких аномальных событий требуется сформировать поведенческое пространство агента. В отличие от классического анализа пользовательской активности, где учитываются действия мыши, клавиатуры или интервалы активности, для ИИ-агентов основными признаками становятся структура траектории, выбор

м

н

о

ж

инструментов, аргументы вызовов, обращения к памяти, ошибки, повторы и соответствие действий исходной задаче. Тогда для каждого шага может быть построен вектор признаков $x_t = (x_{t1}, x_{t2}, \dots, x_{tN}) \in R^N$, а для всей траектории – агрегированный вектор $X_\tau = F(x_1, x_2, \dots, x_k)$, где F – функция агрегации, вычисляющая длину траектории, число вызовов инструментов, долю ошибок, число повторов, количество обращений к памяти, нарушения политик безопасности и другие характеристики. Показатель соответствия действий исходной цели может быть определён как $q_t = sim(g, u_t)$, где g – векторное представление цели задачи, u_t – представление действия агента, $sim(\cdot)$ – мера семантического сходства. Низкое значение q_t при высоком риске инструмента может указывать на отклонение агента от допустимой траектории.

Для выявления нетипичных траекторий может быть использован подход, применяемый при обнаружении аномального поведения пользователей информационных систем [6–10]. Пусть множество наблюдаемых траекторий описывается матрицей признаков $D \in R^{T \times N}$, где T – число наблюдений, N – число признаков. После нормализации признаков автокодировщик обучается восстанавливать нормальное поведение агента: $AE(x) = Decoder(Encoder(x)) = \hat{x}$. Ошибка реконструкции определяется как $\varepsilon(x) = \frac{1}{N}\sum_{j=1}^{N}(x_j - \hat{x}_j)^2$. Если новая траектория плохо восстанавливается автокодировщиком, то есть имеет высокое значение $\varepsilon(x)$, она рассматривается как отклоняющаяся от профиля нормы и потенциально содержащая редкие аномальные события. Использование последовательностных и вероятностных моделей для выявления отклонений согласуется с ранее применявшимися подходами на основе цепей Маркова и скрытых марковских моделей .

Для интерпретации отклонений формируется эталонная матрица поведения. По

распреде[illegible]

И

Сравнение такой подписи с эталонными подписями нормального и аномального поведения позволяет оценить структурное сходство текущей траектории с ранее

Таким образом, задача обнаружения редких аномальных событий, связанных с поведением ИИ-агентов, является развитием классической задачи анализа поведения пользователей информационных систем. В отличие от традиционных подходов анализируется не только итоговое действие, но и вся траектория принятия решений агента: цель, план, выбор инструментов, параметры вызова, наблюдения и изменения состояния.

Предложенный подход позволяет выявлять опасные отклонения, которые могут быть незаметны при контроле только финального результата, и формирует основу для построения систем мониторинга и ограничения поведения автономных ИИ-агентов.

Список литературы:

1. Котенко И. В., Саенко И. Б., Лаута О. С. и др. Атаки и методы защиты в системах машинного обучения: анализ современных исследований // Вопросы кибербезопасности. 2024. № 1(59). С. 24–37.
2. Котенко И. В., Дун Х. Обнаружение атак в Интернете вещей на основе многозадачного обучения и гибридных методов сэмплирования // Вопросы кибербезопасности. 2024. № 2(60). С. 10–21.
3. Liu Y. et al. TrajAD: Trajectory Anomaly Detection for Trustworthy LLM Agents // arXiv. 2026.
4. Advani A. Trajectory Guard: A Lightweight, Sequence-Aware Model for Real-Time Anomaly Detection in Agentic AI // arXiv. 2026.
5. Yao S. et al. ReAct: Synergizing Reasoning and Acting in Language Models // ICLR. 2023.
6. Буйневич М. В., Власов Д. С., Моисеенко Г. Ю. Комбинирование способов выявления инсайдеров больших информационных систем // Вопросы кибербезопасности. 2024. № 3(61). С. 2–13.
7. Буйневич М. В., Моисеенко Г. Ю. Повышение «устойчивости» регламентов деятельности как способ противодействия неумышленному инсайдингу // Вопросы кибербезопасности. 2024. № 6(64). С. 108–116.
8. Ушаков И. А. Обоснование демаскирующих признаков инсайдеров в корпоративных компьютерных сетях // Вестник СПбГУТД. Серия 1. 2024. № 3. С. 95–106.
9. Шелухин О. И., Раковский Д. И. Многозначная классификация компьютерных атак с использованием искусственных нейронных сетей с множественным выходом // Труды учебных заведений связи. 2023. Т. 9. № 4. С. 97–113.
10. Шелухин О. И., Осин А. В., Хализев К. А. Обнаружение аномальной активности пользователей информационной системы на основе анализа кластерных соседств // Методы и технические средства обеспечения безопасности информации. – 2025. – № 34. – С. 11-14.

Шелухин О.И., Рыбаков С.Ю.

Московский технический университет связи и информатики (МТУСИ), Москва

ПОВЫШЕНИЕ ЭФФЕКТИВНОСТИ МНОГОКЛАССОВОЙ КЛАССИФИКАЦИИ КОМПЬЮТЕРНЫХ АТАК МЕТОДАМИ МУЛЬТИФРАКТАЛЬНОГО АНАЛИЗА

Публикация выполнена в рамках гранта на реализацию отраслевой научно-педагогической школы МТУСИ "Современные технологии исследования аномалий в информационной безопасности" по проекту "Обнаружение и прогнозирование редких аномальных событий для обеспечения информационной безопасности" (Пр. 93-х от

Современные компьютерные сети (КС) подвержены воздействию разнообразных компьютерных атак (КА), масштабы и сложность которых непрерывно растут. Традиционные сигнатурные методы обнаружения вторжений КА демонстрируют низкую эффективность против атак нулевого дня и целевых угроз, адаптирующихся к динамике сетевой среды. В связи с этим активно развиваются подходы на основе интеллектуального анализа данных и машинного обучения, которые ориентированы на выявление аномалий, КА и таргетированных атак в трафике КС. Однако и такие классификаторы имеют

следующие недостатки, такие как значительный уровень ложных срабатываний, чувствительность к нестационарности сетевого трафика и, что особенно важно, имеют ограниченную информативность используемых признаков, не способных уловить глубинные структурные изменения трафика компьютерных сетей, вызываемые аномальной активностью и компьютерными атаками.

Перспективным направлением для обнаружения и мониторинга динамики структурных изменений сетевого трафика является использование методов моно- и мультифрактальных характеристик сетевого трафика. Установлено, что нормальный трафик обладает устойчивым свойством самоподобия, тогда как воздействие компьютерных атак приводит к изменению фрактальной размерности и нарушению масштабно-инвариантных свойств. Однако монофрактальные модели не способны адекватно описывать трафик в широком диапазоне временных разрешений, поскольку аномальная активность, как правило, обладает мультифрактальной природой [1,2].

Трафик КС демонстрирует квазистационарное самоподобие в широком диапазоне масштабов, количественно выражаемое через показатель Херста H, связанный с фрактальной размерностью Хаусдорфа соотношением $D_f=2-H$. Для отслеживания структурных изменений в реальном времени оценка H выполняется в скользящем окне, однако прямые вычисления подвержены значительным флуктуациям из-за шумов и нестационарности среды. С целью стабилизации оценок, в [2] предложена модификация вейвлет-разложения на основе жесткой пороговой фильтрации (трешолдинга) детализирующих коэффициентов, что в сочетании с вейвлетом Хаара снижает дисперсию не менее чем на 50% при сохранении несмещенности математического ожидания.

Поскольку показатели самоподобия демонстрируют вариабельность при изменении временного разрешения (в интервале от миллисекунд до минут) при воздействии компьютерной атаки или аномальной активности трансформируется фрактальная природа трафика КС, что выражается в переходе от монофрактальной к более сложной мультифрактальной организации. Для формализации данного явления введено понятие мультифрактального спектра фрактальной размерности, определяемого как последовательность текущих оценок фрактальной размерности $\widehat{H}_{t_{Д_i}}$ в окне анализа ф

и Экспериментальная проверка метода композиции алгоритмов машинного обучения и мультифрактального анализа проводилась на открытом наборе данных, представленном в базе Kitsune [3]. Этот набор содержит размеченный трафик интернета вещей, включающий экземпляры атак «Mirai Botnet», «OS Scan» и нормальную активность. Исходный набор включал 115 атрибутов, сформированных методом инкрементной статистики по пяти временным окнам длительностью от 100 мс до 1 мин и коэффициентам «старения» данных. Атрибутное пространство исходного набора данных было расширено до 120 атрибутов путём добавления пяти новых признаков — текущих оценок H, вычисленных соответственно для каждого из пяти временных масштабов. Эксперименты разделялись на два этапа:

- бинарная классификация - нормальный трафик (Normal) vs атака «Mirai Botnet».
- многоклассовая классификация - нормальный трафик (Normal) vs атаки «Mirai Botnet» и «OS Scan».

В качестве базового алгоритма машинного обучения выбран ансамблевый метод «Случайный лес» (Random Forest, RF). Тестирование проводилось при глубине решающего дерева depth ∈ {2,5}. Верификация качества классификации проводилась с использованием набора стандартных метрик таких как precision (точность обнаружения), recall (полнота выявления), F1-меры, и интегрального показателя ROC_AUC_OVR, характеризующего площадь под ROC-кривой при бинаризации многоклассовой задачи по принципу «один против всех». Чтобы гарантировать достоверность оценки значимости новых атрибутов, во всех записях, не принадлежащих к классу атаки Mirai, атрибуты МСФР инициализировались нулевыми значениями.

Результаты бинарной классификации демонстрируют, что при добавлении атрибутов МСФР метрики точности, полноты и F1-меры достигают единичных значений независимо от конфигурации глубины дерева, что свидетельствует о полной разделимости классов в расширенном признаковом пространстве.

В задаче многоклассовой классификации при *depth=2 (рис.1)* без учета МСФР эффективность распознавания атаки Mirai относительно OS Scan и Normal составляет 7,6%), при этом классификация атаки OS Scan остается стабильно высокой (*AUC≈0,99*). Увеличение глубины дерева до *depth=5* в сочетании с расширенным признаковым пространством обеспечивает достижение метрик, близких к идеальным (*AUC≈1,0* по всем классам).

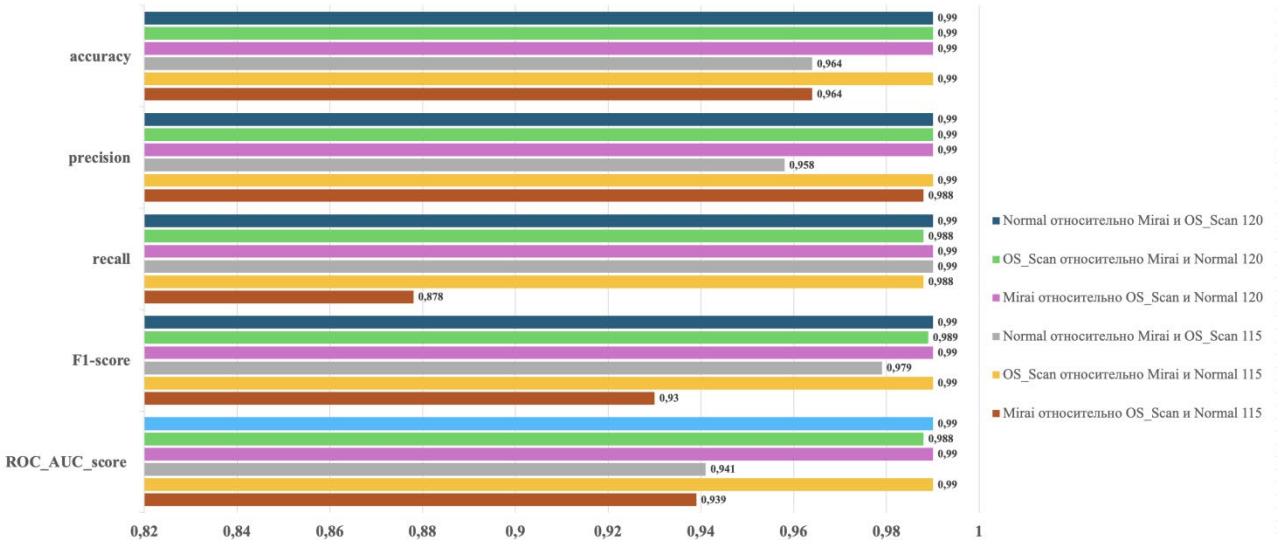


Рис. 1. Визуализация эффективности многоклассовой классификации алгоритмом RF

Композиция алгоритмов мультифрактального анализа и машинного обучения демонстрирует высокую эффективность в задаче многоклассовой классификации КА. Расширение исходного набора атрибутов параметрами мультифрактального спектра позволяет повысить точность распознавания на 7–10% и существенно сократить число ошибок первого рода. Полученные экспериментальные данные обосновывают применение характеристик МСФР как устойчивых инвариантов для мониторинга структурных изменений в динамике сетевого трафика КС и детектирования аномальной активности в условиях нестационарности сетевого трафика.

Список литературы:

.И. Шелухин, С.Ю. Рыбаков, А.В. Ванюшина. Оценка характеристик мультифрактального спектра фрактальной размерности сетевого трафика и компьютерных атак в IoT // Труды учебных заведений связи. – 2024. – Т. 10, № 3. – С. 104–115.

2. O.I. Sheluhin, S.Y. Rybakov. Characteristics Assessment of Multifractal Spectrum of Fractal Dimension IoT-Traffic // Systems of Signal Synchronization, Generating and Processing in Telecommunications (SYNCHROINFO). – 2024. – P. 1–6.
3. Alabdulatif A., Rizvi S. Machine Learning Approach for Improvement in Kitsune NID // Intelligent Automation & Soft Computing. 2022. Т. 32. С. 827-840. DOI: 10.32604/iasc.2022.021879.

Шутков А.С.
ПФ АО «НТЦ «Ат лас», Пенза

АВТОМАТИЗИРОВАННЫЕ СРЕДСТВА ВЕРИФИКАЦИИ ЗАЩИЩЕННОСТИ КОНФИДЕНЦИАЛЬНОЙ ИНФОРМАЦИИ В ОБЕЗЛИЧЕННЫХ НАБОРАХ ДАННЫХ

В настоящее время все операторы сотовой связи и разнообразные сервисы собирают и анализируют большие объемы пользовательских данных, содержащих конфиденциальную информацию, в том числе персональные данные. Защита конфиденциальной информации участников базы (абонентов сотовой связи, пользователей сервисов) в больших базах данных является актуальной задачей. Защита информации при хранении может быть обеспечена традиционным шифрованием. Защита информации при обработке (статистическом анализе) является более сложной задачей.

Для статистического анализа совокупности записей в базе данных (выборки) аналитик, используя различные запросы к базе, вычисляет значения статистик – функций, зависящих от реальных данных. Цель защищенной обработки данных состоит в организации ответов на запросы таким образом, чтобы аналитик мог изучать свойства совокупности записей в целом, но не имел возможности восстановить конфиденциальную информацию (персональные данные) отдельных участников базы данных. Задачу сохранения конфиденциальности персональных данных (индивидуальных записей) при статистической обработке информации, содержащейся в базе данных, можно определить, как задачу обезличивания.

Существуют две различные модели обезличивания: неинтерактивная и интерактивная. В неинтерактивной модели собранные данные перед использованием «очищаются»: персональные данные искажаются или удаляются. Неинтерактивные методы обеспечивают ограниченную защиту, позволяя восстанавливать данные субъектов при наличии дополнительной информации. В интерактивной модели обезличивание состоит в использовании защищенного интерфейса, с помощью которого аналитик может создавать запросы о данных и получать (возможно, искаженные) ответы. В этом случае говорят о статистическом обезличивании и рассматривают понятие дифференциальной защиты.

Для оценки качества методов обезличивания рассматриваются различные теоретические оценки сложности и точности восстановления конфиденциальных персональных данных. Однако теоретические оценки носят общий характер и не могут учесть специфику реальных данных.

Для подтверждения защищенности реальных конфиденциальных персональных данных с применением конкретных методов обезличивания разработаны различные программные средства верификации.

Одним из инструментов верификации защищенности конфиденциальных данных с применением неинтерактивных методов обезличивания является *pyCANON* – Python библиотека, предназначенная для оценки защищенности с помощью наиболее распространенных методов анонимизации: k-анонимность, (α,k)-анонимность, l-разнообразие, энтропия l-разнообразия, рекурсивное (c,l)-разнообразие, t-близость, базовое и расширенное β-подобие, δ-нарушение конфиденциальности. Основное преимущество этой библиотеки заключается в получении полного отчета о параметрах, определяющих устойчивость методов обезличивания данных к различным атакам, при которых обеспечивается защищенность конфиденциальных данных.

Для верификации защищенности конфиденциальных данных с применением интерактивных методов обезличивания (дифференциальной защиты) предлагается инструмент DiPC (Differential Privacy Checker), представляющий собой процедуру принятия решений. Он относится к классу вероятностных while программ, называемых DiPWhile, для которых задача проверки дифференциальной защиты является разрешимой.

Создание универсального средства верификации защищенности конфиденциальных данных с применением интерактивных и неинтерактивных методов обезличивания является следующим шагом, который помог бы продвинуться в обосновании защищенности конфиденциальной информации в обезличенных наборах данных.

Югай П.Э., Москвин Д.А.
Санкт-Петербургский политехнический университет Петра Великого, Санкт-Петербург

ИСПОЛЬЗОВАНИЕ ДИНАМИЧЕСКОГО ИНДЕКСА НЕСОГЛАСИЯ АНСАМБЛЯ МОДЕЛЕЙ МАШИННОГО ОБУЧЕНИЯ ВО ВРЕМЕННЫХ РЯДАХ ДЛЯ ОБНАРУЖЕНИЯ СОСТЯЗАТЕЛЬНЫХ АТАК В СЕТЕВЫХ ДАННЫХ

Состязательные атаки на алгоритмы машинного обучения (МО) представляют собой серьезную угрозу для современных систем защиты от сетевых атак, где модели используются для автоматического обнаружения вторжений, классификации трафика и идентификации вредоносного программного обеспечения (ПО) [1]. Данные атаки основаны на преднамеренном создании минимальных возмущений в сетевых данных, которые приводят к некорректной работе моделей МО, а именно к возникновению ложноотрицательных и ложноположительных результатов [2].

Существует ряд актуальных методов и подходов защиты моделей машинного обучения в системах обнаружения и предотвращения вторжений от состязательных атак. Классическим и распространенным методом является состязательное обучение, в котором модель МО обучается на легитимных и состязательных примерах, генерируемых в процессе обучения с помощью различных известных алгоритмов. Данный метод повышает точность обнаружения известных состязательных атак, однако модель остается уязвимой к атакам, не представленным в обучающей выборке [3]. Следующим подходом является предобработка входных данных с помощью детерминированных или стохастических преобразований: сжатие признаков, квантование, добавление шума. В данном случае отсутствует необходимость в модификации архитектуры модели или процесса обучения, однако такой подход уязвим к адаптивным атакам, а также существует высокий риск снижения точности выявления легитимных данных моделью МО [4]. Иным подходом является оборонительная дистилляция, при которой знания из первичной модели-учителя передаются модели-ученику с использованием сглаженной функции Softmax с повышенным температурным параметром. Этот подход позволяет маскировать градиенты модели МО и снижать ее чувствительность к малым состязательным искажениям сетевых признаков, однако он не защищает от атак «черного ящика», атак с суррогатной моделью и оптимизационных алгоритмов, способных обойти сглаженный ландшафт функции потерь [5].

В связи с этим существует необходимость в разработке подходов для выявления состязательных атак в системах защиты от сетевых угроз, которая обусловлена фундаментальными ограничениями существующих методов, в частности их низкой обобщающей способностью к ранее неизвестным типам возмущений и уязвимостью к адаптивным атакам, целенаправленно обходящими механизмы защиты через учет специфики модели МО или использование суррогатных моделей.

В современных системах защиты используются ансамбли моделей МО для обнаружения сетевых угроз [6]. Для системы защиты от сетевых угроз предлагается определенный ансамбль моделей МО, а также вводится метрика – динамический индекс несогласия, который является показателем динамики изменения результатов работы алгоритмов машинного обучения во временном промежутке. Динамический индекс несогласия и сформированный ансамбль методов МО являются основой для формирования временных рядов и последующего их анализа с целью выявления состязательных атак.

В данной работе представлен сравнительный анализ алгоритмов машинного обучения для анализа временных рядов динамического индекса несогласия. Обоснован выбор двухуровневой архитектуры LSTM-модели для выявления состязательных атак. Предложен механизм обнаружения состязательных атак, основывающийся на анализе временных рядов для краткосрочных и долгосрочных последовательностей динамических индексов несогласия ансамбля моделей машинного обучения. Продемонстрирована работоспособность предложенного механизма обнаружения состязательных атак.

Библиографический список:

1. Kulikov D.A., Platonov V.V. Adversarial attacks on intrusion detection systems using an LSTM classifier // Problems of information security. Software security. – 2021. – No. 2. – P. 48-56.
2. Ivanova O. D., Kalinin M. O. Hybrid method for identifying evasion attacks aimed at intrusion detection systems // Problems of information security. Software security. – 2023. – No. 1. – P. 104-110.
3. Wu B., Wei S., Zhu M., Zheng M., Zhu Z., Zhang M., Chen H., Yuan D., Liu L., Liu Q. Defenses in Adversarial Machine Learning: A Survey. 2023. URL: https://arxiv.org/pdf/2312.08890.
4. Ennaji S., De Gaspari F., Hitaj D., Bidi A. K., Mancini L. V. Adversarial Challenges in Network Intrusion Detection Systems: Research Insights and Future Prospects. 2024. URL: https://arxiv.org/html/2409.18736v3.
5. Zhang J., Nikolić K., Carlini N., Tramèr F. Gradient Masking All-at-Once: Ensemble Everything Everywhere Is Not Robust. 2024. URL: https://arxiv.org/html/2411.14834v1.
6. Arreche O., Bibers I., Abdallah M. A Two-Level Ensemble Learning Framework for Enhancing Network Intrusion Detection Systems // IEEE. 2024. URL: https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=10540382.

2. КИБЕРУСТОЙЧИВОСТЬ КРИТИЧЕСКОЙ ИНФОРМАЦИОННОЙ ИНФРАСТРУКТУРЫ

Альшанская Т.В.
Самарский университет, Самара

ОСОБЕННОСТИ БЕЗОПАСНОЙ РАЗРАБОТКИ ПО: СОВРЕМЕННЫЕ ПОДХОДЫ

В условиях динамичного роста и сложности кибератак процессы безопасной разработки программного обеспечения (ПО) должны выстраиваться на основе систематизации требований нормативно-правовых актов (НПА), регламентов, методологий, фреймворков, а также на основе применения соответствующих инструментов. Это позволяет минимизировать риски информационной безопасности и усовершенствовать качество разработки программного обеспечения. Наличие уязвимости в коде, не корректная настройка инфраструктуры, недостаточная защита информации и данных могут способствовать утечкам, различным вариантам негативных последствий.

Комплексный подход обеспечения безопасной разработки ПО (РБПО) включает интеграцию необходимых инструментов на всех этапах жизненного цикла программного обеспечения (SDLC), а также проведение аудита безопасности процессов разработки ПО с учетом всех требований нормативно-правовой базы.

Данный подход реализуется на основе выполнения требований стандарта, принятых в 2024 году и ранее, указанных в таблице 1. Необходимо также применение нормативных документов ФСТЭК: приказ от 01.12.2023 № 240, Методика выявления уязвимостей и недекларированных возможностей в программном обеспечении. Процессы разработки проекта на всех этапах (SDLC) должны базироваться на создании основного документа «Руководство по безопасной разработке программного обеспечения», с учетом требований приказа № 240, ГОСТ Р 56939-2024. [1]

Таблица 1. НПА безопасной разработки ПО

Документ	Наименование
ГОСТ Р 56939–2024	Защита информации. Разработка безопасного программного обеспечения. Общие требования
ГОСТ Р 71206–2024	Защита информации. Разработка безопасного программного обеспечения. Безопасный компилятор языков С/С++. Общие требования
ГОСТ Р 71207–2024	Защита информации. Разработка безопасного программного обеспечения. Статический анализ программного обеспечения. Общие требования
ГОСТ Р 56920-2024	Системная и программная инженерия. Тестирование программного обеспечения. Общие положения
ГОСТ Р 58412-2019	Защита информации. Разработка безопасного программного обеспечения. Угрозы безопасности информации при разработке программного обеспечения

На каждом этапе жизненного цикла ПО: проектирование, кодирование, тестирование развертывание, поддержка и сопровождение важно применение соответствующих инструментов, обеспечивающих качественную разработку. На начальном этапе происходит либо формирование команды DevSecOps, либо четкое распределение ролей в уже существующей команде РБПО, затем дальнейшая экспертиза. Применение методологии DevSecOps создает возможность непрерывной интеграции безопасности в каждый этап создания ПО. [2]

На рисунке 1 показаны примеры возможных инструментов на разных этапах жизненного цикла разработки ПО.

Рисунок 1- Инструменты безопасной разработки ПО на разных этапах жизненного цикла

Так же существует ряд инструментов для управления уязвимостями (Vulnerability Management Tools) и угрозами (таблица 2).

Таблица 2 – Инструменты управления угрозами и уязвимостями

Управление уязвимостями		**Управление угрозами**	
Nessus	Сканер, поддерживает множество платформ и может обнаруживать тысячи различных типов уязвимостей	MITRE ATT&CK	более 200 техник, используемых атакующими, и может помочь разработчикам и администраторам разрабатывать стратегии защиты
OpenVAS	Открытый сканер (более 50 тысяч проверок), поддерживает множество платформ и может обнаруживать уязвимости в операционных системах, сетях и веб-приложениях	ThreatConnect	Позволяет идентифицировать, классифицировать и митигировать угрозы; интегрируется с другими инструментами
Vulners	Инструмент для управления уязвимостями, интегрируется со многими другими инструментами для безопасности, такими как Nessus, OpenVAS и OWASP ZAP	OWASP	Набор критичных рисков WEB-приложений, включая угрозы нарушения контроля доступа, криптографические ошибки

Указанные в таблице инструменты позволяют своевременно выявлять угрозы и реагировать на различные инциденты до внедрения ПО на готовом объекте. Системная организация процессов разработки способствует минимизации различных рисовых ситуаций, сокращение статей затрат на устранение уязвимостей и последствий атак. Таким образом, РБПО представляет комплексный процесс на основе соответствующего инструментария, обеспечивающего разработчиков надежными средствами, методологиями, фреймворками для создания, защищённого ПО.

Библиографический список:

1. Стандартизация безопасной разработки: теория и практика/ URL: https://www.angarasecurity.ru/stati/standartizatsiya-bezopasnoy-razrabotki-teoriya-i-raktika/(2025).
2. Разработка безопасного программного обеспечения: DevSecOps как обязательный стандарт:https://infars.ru/blog/razrabotka-bezopasnogo-programmnogo-obespecheniya_-devsecops-kak-obyazatelnyy-standart/ (2024).

Васильев О.С., Крундышев В.М.
Санкт-Петербургский политехнический университет Петра Великого, Санкт-Петербург

ФРЕЙМВОРК ДЛЯ ОЦЕНКИ УСТОЙЧИВОСТИ СИСТЕМ РАСПРЕДЕЛЕННОГО РЕЕСТРА «УМНОГО ГОРОДА» К АТАКАМ ДЕАНОНИМИЗАЦИИ

Исследование выполнено за счет гранта Российского научного фонда №24-11-20005, https://rscf.ru/project/24-11-20005/; грант Санкт-Петербургского научного фонда (договор №24-11-20005 о предоставлении регионального гранта)

Цифровизация городов является одним из приоритетных направлений последних лет. Ожидаемый эффект от цифровизации заключается в повышении безопасности жителей мегаполисов, росте качества предоставляемых услуг и дополнительном импульсе для развития экономики [1]. Технологическую основу цифровизации составляют беспроводные сети связи нового поколения, устройства Интернета вещей, а также системы распределенного реестра, широко используемые при создании платежных систем (например, для оплаты проезда по платным магистралям и парковки автомобилей в соответствующих зонах). Однако, внедрение новых технологий неразрывно связано с рисками. Одной из основных проблем возникающих при внедрении систем распределенного реестра становятся атаки деанонимизации в отношении пользователей городской инфраструктуры. Анализ смежных работ в данной области показал, что злоумышленники используют методы статистического анализа, теорию графов и специализированные утилиты для отслеживания движения денежных средств по сети [2, 3]. Для проверки научных гипотез и тестирования прототипов в области создания механизмов защиты, как правило, используются симуляторы и фреймворки. Однако, на сегодняшний день существующие решения, предназначенные для моделирования распределенных реестров «умных городов», имеют недостатки, ограничивающие их использование при оценке эффективности разрабатываемых методов защиты и протоколов. Целью данного исследования является создание фреймворка, позволяющего оценить устойчивость систем распределенного реестра к атакам деанонимизации с учетом особенностей функционирования цифровых инфраструктур «умных городов».

В рамках данного исследования разработан фреймворк, состоящий из генератора транзакций, анализатора для моделирования атак деанонимизации, а также протоколы, предназначенные для защиты от таких атак. Кроме того, реализован модуль визуализации для отображения графа транзакций на основе данных, извлеченных из заданного диапазона блоков. Эта подсистема обеспечивает: извлечение транзакций и связей между ними с использованием модели UTXO; построение направленного графа транзакций; наложение эталонной и обнаруженной цепочек; и интерактивную визуализацию графа с возможностью фильтрации, и уменьшения визуального шума. Выходные данные модуля представляют собой HTML-файл с интерактивным графом транзакций, предназначенным для визуального анализа и проверки результатов обнаружения. Разработанный фреймворк реализован в виде набора модулей Python, взаимодействующих друг с другом и с узлом Bitcoin Core 30.0 через RPC-интерфейс. Архитектура генератора транзакций является модульной, что позволяет разделить функциональность на логические компоненты и упростить сопровождение и модификацию кода. Модуль моделирования атак деанонимизации поддерживает два режима работы: на основе анализа сумм и на основе временного и потокового анализа.

Для оценки устойчивости систем распределенного реестра к атакам деанонимизации в локальной среде блокчейна Bitcoin был реализован пртокол CoinJoin, предполагающий совместное формирование транзакции несколькими независимыми участниками, в котором входные данные из разных кошельков объединяются в одну транзакцию, а выходные данные имеют одинаковый номинал, что усложняет установление соответствия между входными и выходными данными транзакции. Алгоритм генерации транзакций CoinJoin включает следующие шаги: выбор набора участников CoinJoin; формирование структуры транзакции с эквивалентными выходными данными; распределение комиссий между участниками; совместное подписание транзакции с использованием PSBT; обновление отслеживаемого UTXO. На каждом шаге генерации определяется основной кошелек, которому принадлежит текущий отслеживаемый UTXO. Эксперименты показали, что использование транзакций CoinJoin значительно усложняет задачу деанонимизации и приводит к снижению точности анализаторов. На рис. 1 показан результат наложения восстановленной цепочки (синим цветом) на исходную цепочку (красным цветом); в

случае успешного наложения вершины выделены фиолетовым цветом. График показывает, что злоумышленник не смог идентифицировать все транзакции из исходной цепочки, а также обнаружил некоторые другие транзакции, которые не являлись частью исходной цепочки.

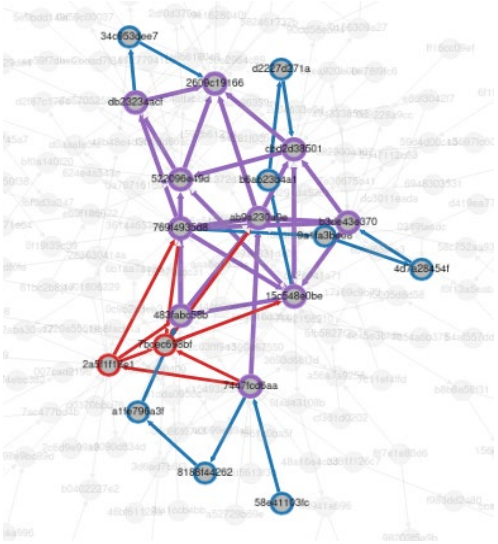


Рис. 1. Результат работы модуля визуализации

Таким образом, разработанный фреймворк позволил убедиться, что применение протокола CoinJoin позволяет повысить устойчивость к атакам деанонимизации на основе методов анализа сумм, а также временного и потокового анализа. План дальнейших работ включает в себя интеграцию разработанного фреймворка в сетевой эмулятор NS-3, предназначенный для моделирования беспроводных динамических сетей. Такое сочетание позволит повысить точность и полноту моделирования, в результате чего оценка устойчивости систем распределенного реестра «умного города» к атакам деанонимизации станет более объективной.

Список литературы:

1. Калинин М. О. Обеспечение безопасности технологии блокчейн в экосистемах Цифровой экономики / М. О. Калинин, А. С. Коноплев // Стратегическое управление цифровой трансформацией интеллектуальной экономики и промышленности в новой реальности : Монография. – Санкт-Петербург : ПОЛИТЕХ-ПРЕСС, 2024. – С. 396-421.
2. Петренко С. А. Модель квантовых угроз безопасности информации для национальных блокчейн-экосистем и платформ / С. А. Петренко, А. А. Балябин // Вопросы кибербезопасности. – 2025. – № 1(65). – С. 7-17.
3. Охлопков Н. М. Анонимизация сетевого трафика на основе чесночной маршрутизации в защищенной блокчейн-инфраструктуре умного города / Н. М. Охлопков, М. О. Калинин // Неделя науки ИКНК : Материалы Всероссийской научно-практической конференции, Санкт-Петербург, 01–20 апреля 2026 года. – Санкт-Петербург: ПОЛИТЕХ-ПРЕСС, 2026. – С. 153-155.

Зубков Е.А., Овасапян Т.Д., Зегжда Д.П.
Санкт-Петербургский политехнический университет Петра Великого, Санкт-Петербург

МЕТОД ОЦЕНКИ УСТОЙЧИВОСТИ КФС, ОСНОВАННЫЙ НА МЕТОДИКЕ ОЦЕНКИ ЗАЩИЩЕННОСТИ ПО, В ОТНОШЕНИИ ЦЕЛЕНАПРАВЛЕННЫХ АТАК

Одно из ключевых свойств киберфизических систем (КФС) – свойство устойчивости, – показывает, в какой степени КФС способны реализовывать свое функциональное назначение и самовосстанавливаться в случае, когда в их отношении выполняется реализация целенаправленных кибератак.

Наибольший риск для КФС составляют целенаправленные атаки, связанные с компрометацией закрытой инфраструктуры, в результате которых злоумышленник получает несанкционированный доступ к программно-аппаратным ресурсам КФС и закрытым данным (рисунок 1).

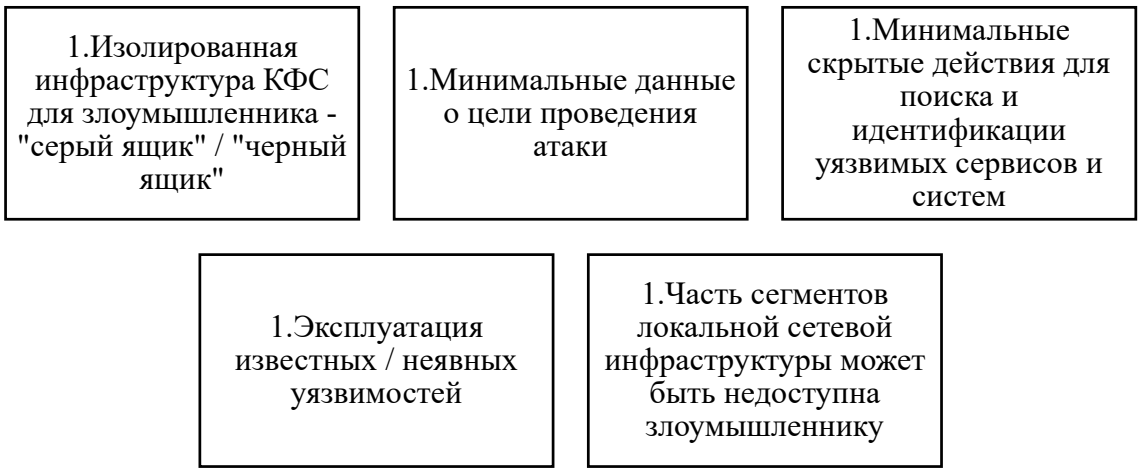


Рисунок 1 – Особенности проведения целенаправленных атак в отношении КФС

Для того, чтобы оценить степень устойчивости КФС в отношении обозначенных целенаправленных атак, в рамках данной работы был разработан метод оценки устойчивости, основанный на следующих данных:

1. Результаты методики оценки защищенности ПО $s_i \in S_{h_j}$, используемого на объектах КФС $h_j \in H$. Предполагается, что в отношении каждого экземпляра используемого ПО справедливо наличие как известных уязвимостей V_{s_i}, так и неявных уязвимостей V'_{s_i}.
2. Результаты оценки действий злоумышленника над объектами КФС (H) в рамках целенаправленной атаки.

Злоумышленник, получивший начальный доступ к закрытой инфраструктуре КФС, выполняет ряд действий для реализации целенаправленной атаки:

- выявление и идентификация уязвимых сетевых сервисов;
- эксплуатация известных / неявных уязвимостей;
- захват соседних объектов КФС в случае успешной эксплуатации уязвимостей (с целью поэтапного продвижения к цели атаки).

В свою очередь, средства обеспечения кибербезопасности КФС, в случае детектирования вредоносной активности, предпринимают определенные действия для прекращения атаки:

- изоляция зараженных объектов КФС;
- ограничение доступа к критической инфраструктуре.

Следовательно, устойчивость КФС напрямую зависит от того, какие действия выполняются злоумышленником в рамках цепочки атаки и как средства защиты способны ограничить цепочку атаки.

Для оценки действий злоумышленника в КФС применяется математическая модель (рисунок 2), основанная на представлении КФС в виде общей информационно-графовой модели $G =< V, E >$ и пошаговой оценке действий злоумышленника с использованием:

– параметров программного / аппаратного объекта КФС (функциональная значимость $T(h_i)$, совокупная уязвимость $V_{Total}(h_i)$ и прочее);

– особенностей поведения злоумышленника (оценка привлекательности следующего атакуемого объекта КФС $A(h_i)$, оценка риска обнаружения в рамках эксплуатации $R_{Detection}(h_i)$, оценка стоимости эксплуатации известных / неявных уязвимостей $C_{Exploitation}(h_i)$ и прочее).

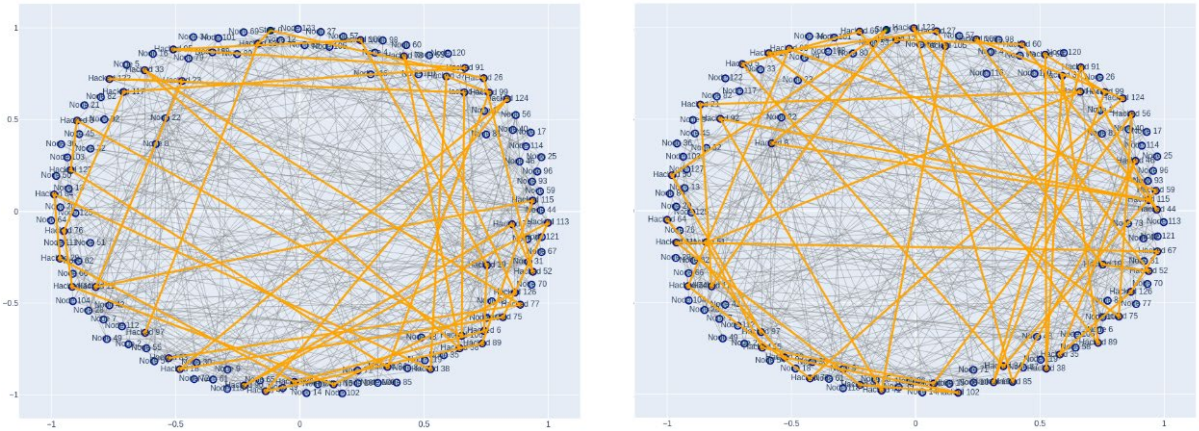


Рисунок 2 – Визуализация действий злоумышленника в инфраструктуре КФС (слева – без учета неявных уязвимостей, справа – с учетом неявных уязвимостей)

Кессаринский Л.Н.[1], Давыдов Г.Г.[1], Дураковский А.П.[1], Сидорин Ю.Ю.[2]
[1]НИЯУ МИФИ, Москва
[2]АНО «Доверенная платформа», Москва

МЕТОД ФОРМИРОВАНИЯ ТРЕБОВАНИЙ ПО ЗАЩИТЕ ОТ АТАК НА ПРОЦЕССЫ СТАДИЙ ЖИЗНЕННОГО ЦИКЛА СОВРЕМЕННЫХ ЦИФРОВЫХ СВЕРХБОЛЬШИХ ИНТЕГРАЛЬНЫХ СХЕМ

Цифровые сверхбольшие интегральные схемы (СБИС) являются основой современной критической информационной инфраструктуры, в тоже время, действующие документы стандартизации гражданской микроэлектроники были разработаны в прошлом веке и описывают преимущественно предприятия полного цикла (сам разрабатывает, сам производит, сам поставляет), что справедливо лишь для иностранных изделий военной или космической категорий качества, а также для отечественной продукции тех.процесса 0,5 мкм и больше.

Экономика жизненного цикла современных СБИС приводит к усложнению производственно-сбытовых цепочек: широкой кооперации, сложных логистических связей, узкой специализации каждого участника (который часто не может перепроверить результат предыдущего процесса и вынужден верить документам), а также высокой сложности самих изделий. Микросхему из 1 млрд. транзисторов просто невозможно протестировать в 100% состояний, режимов и условий работы за разумное время (меньше года). Единственным

рациональным решением проблемы обеспечения доверия к СБИС становится контроль процессов стадий жизненного цикла (ПСЖЦ) - потребитель или уполномоченный им орган вынужден проводить контроль или экспертизу (постфактум) процессов, представленных на рис. 1., а именно: проверять как микросхема была разработана, как изготовлена, как хранилась, как поставлялась - анализировать полную и достоверную информацию о СБИС, от идеи создания до «ворот» своего предприятия.

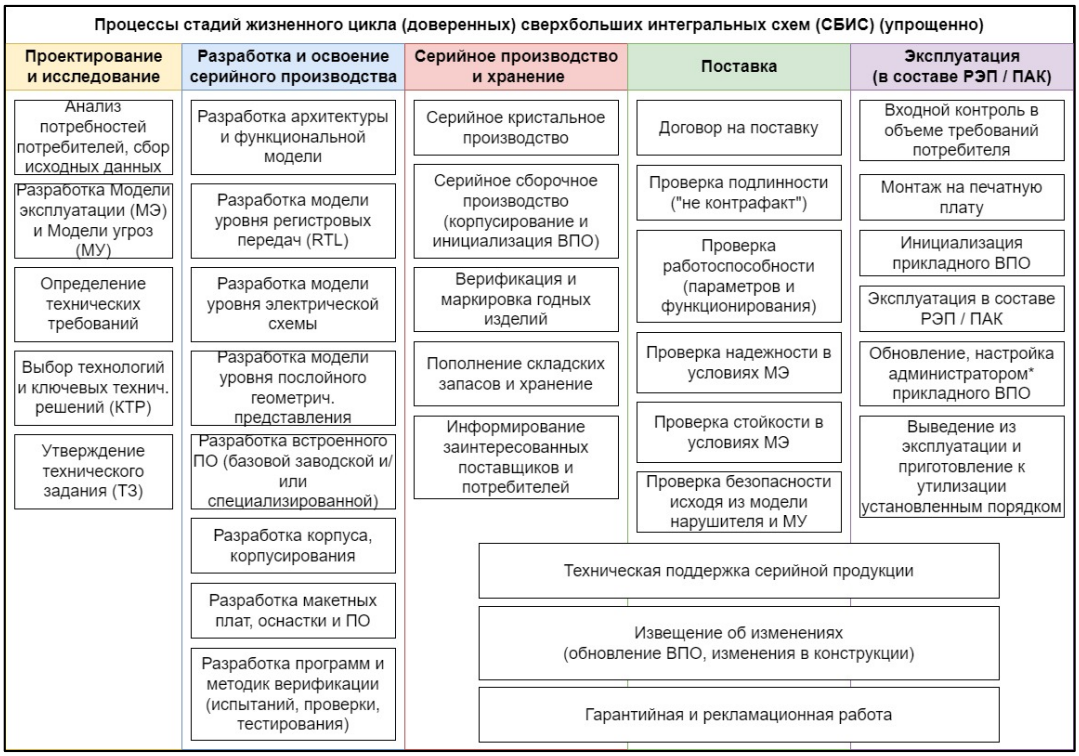


Рис. 1. Упрощенная схема процессов стадий жизненного цикла СБИС

В тоже время, аналогичную задачу решает злоумышленник - чтобы иметь возможность эксплуатировать уязвимость или закладку в СБИС при эксплуатации в составе атакуемой аппаратуры, ее [закладку] следует заложить на предыдущих стадиях жизненного цикла.

Таким образом, поверхность атаки для современных СБИС расширяется на ПСЖЦ т.е. на инфраструктуру предприятий разработки, производства, поставки, а также непосредственно на элементы ПСЖЦ («активы»): документацию; компоненты, полуфабрикаты и материалы; технологическое, измерительное и испытательное оборудование; программные инструменты - САПР, библиотеки сложно-функциональных блоков; оказать воздействие на персонал. Соответственно и меры обеспечения доверия к СБИС трансформируются в меры защиты от угроз атак на ПСЖЦ.

Авторами предлагается метод формирования требований мер защиты от атак на ПСЖЦ СБИС СИОД, который состоит из следующих этапов:

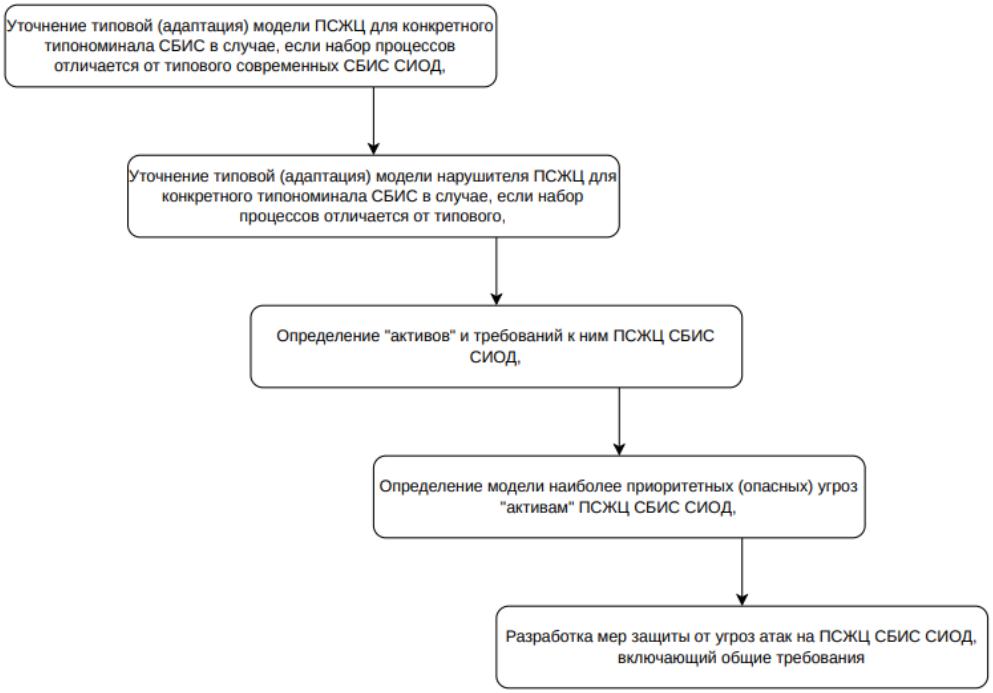


Рис. 2. Схема метода разработки требований мер защиты от атак на ПСЖЦ

Результатом выполнения предложенного метода должен стать список «активов» всех процессов, список угроз данным «активам», прошедших приоритезацию, а также перечень мер защиты. Довольно часто получается, что стандартные меры защиты работают для разных угроз: контроль версий, целостности и подлинности документации, ролевой доступ персонала к ресурсам, журналирование работы оборудования (режимов работы, отчетов испытаний) и действий сотрудников.

Список литературы:

[1] Кессаринский, Л. Н. Подход к заданию общих требований к доверенной электронной компонентной базе для регулируемого рынка критической информационной инфраструктуры в вопросах и ответах / Л. Н. Кессаринский, А. Ю. Никифоров // Безопасность информационных технологий. – 2025. – Т. 32, № 1. – С. 8-16. – EDN EENLAL.

[2] Кессаринский, Л. Н. Участие в 34-ой конференции «Методы и технические средства обеспечения безопасности информации 2025» (МиТСОБИ 2025) с докладом «подход к формированию модели угроз доверенных интегральных микросхем для доверенных Пак» (Кессаринский Л.Н., Сидорин Ю.Ю., Никифоров А.Ю., Дураковский А.П.) / Л. Н. Кессаринский // Безопасность информационных технологий. – 2025. – Т. 32, № 3. – С. 177-179. – DOI 10.26583/bit.2025.3.15. – EDN ZXVVNY.

Мохов Е.О., Лаврова Д.С.

СИСТЕМАТИЗАЦИЯ МЕТОДОВ ОЦЕНКИ КИБЕРУСТОЙЧИВОСТИ И ИХ ПРИМЕНЕНИЕ В IoT-СИСТЕМАХ

Развитие теории защиты информации в последнее десятилетие характеризуется смещением акцента с парадигмы предотвращения инцидентов на парадигму киберустойчивости – способности системы предвидеть неблагоприятные воздействия, выдерживать их, восстанавливаться после компрометации и адаптироваться с учётом

полученного опыта [1]. Особую актуальность данный переход приобретает применительно к системам интернета вещей (Internet of Things, IoT), где ограниченные вычислительные ресурсы оконечных устройств, гетерогенность протоколов, физическая доступность узлов и продолжительный жизненный цикл (10–20 лет в промышленном сегменте) делают полное предотвращение компрометации недостижимым.

Несмотря на наличие развитой методической базы оценки киберустойчивости, большинство существующих подходов разрабатывались для корпоративных информационных сред и при прямом переносе на IoT демонстрируют ограниченную применимость. Это обусловливает актуальность задачи систематизации методов и формирования рекомендаций по их выбору в зависимости от класса защищаемой IoT-системы.

В настоящей работе проведён обзор методов оценки киберустойчивости, по итогам которого они классифицированы на четыре группы: фреймворки соответствия, количественное моделирование рисков, моделирование угроз и динамические методы. К первой группе отнесены NIST Cybersecurity Framework версии 2.0 [1], отраслевой стандарт IEC 62443 для промышленных систем автоматизации [2] и матрица тактик и техник MITRE ATT&CK for ICS [3]. Ко второй – методология количественного анализа рисков FAIR [4], графы атак с расчётом метрик кратчайшего пути и вероятности компрометации, а также марковские модели динамики состояний системы, обеспечивающие оценку среднего времени до компрометации (MTTC) и до восстановления (MTTR). К третьей – методики моделирования угроз STRIDE [5] и PASTA [6]. К четвёртой – динамические индексы киберустойчивости (Cyber Resilience Index, CRI) и методы машинного обучения для непрерывного обнаружения аномалий [7].

Анализ показал, что универсального метода, применимого ко всему многообразию IoT-систем, не существует, что обусловливает необходимость классификации последних для целей выбора адекватного подхода. В работе предложено разграничение IoT-систем по отраслевому домену и уровню критичности с выделением семи классов: домашних (HIoT), промышленных (IIoT), медицинских (IoMT), транспортных (V2X/IoV), городских (Smart City), энергетических и сельскохозяйственных. Для каждого класса определён ключевой атрибут защищённости: для HIoT – приватность и безопасность пользователей, для IIoT – непрерывность технологического процесса с учётом требований функциональной безопасности (IEC 61508), для IoMT – жизнеобеспечение и соответствие нормативным требованиям [8], для V2X – обеспечение работы в реальном времени и противодействие подмене сообщений и спуфингу спутниковой навигации, для городских – устойчивость к масштабным распределённым воздействиям с учётом публичного характера сервисов городского управления, для энергетических систем – доступность и противодействие каскадным отказам объектов критической информационной инфраструктуры, для сельскохозяйственных – автономное функционирование в условиях нестабильной связи и физическая защищённость удалённо размещенных узлов.

На основе сопоставления требований каждого класса с характеристиками методов сформирована матрица рекомендаций. Для систем критической инфраструктуры (IIoT, энергетика) обоснован комплексный подход, объединяющий стандарт IEC 62443 как структурную основу, матрицу ATT&CK for ICS для оценки покрытия защитных мер и графы атак для верификации сетевой сегментации согласно модели Purdue. Для медицинских систем целесообразно применение STRIDE на этапе разработки, FAIR – для обоснования компенсирующих контрольных мер в условиях невозможности оперативного устранения уязвимостей, и непрерывного мониторинга на основе облегчённых моделей машинного обучения на этапе эксплуатации. Для городских систем рекомендован гибридный подход на основе соответствующих матриц ATT&CK, графов атак и методологии FAIR. Для транспортных систем рекомендован переход от периодических аудитов к непрерывному мониторингу на основе динамических CRI и моделей машинного обучения. Для домашних IoT-систем достаточным признано применение базового профиля

NIST CSF в сочетании с облегчёнными моделями обнаружения аномалий, развёртываемыми на уровне шлюзовых устройств. Для сельскохозяйственных систем – аналогичный базовый профиль NIST CSF, дополненный графами атак для оценки рисков физического доступа к удалённо размещённым узлам [9].

Полученные результаты позволяют сформировать обоснованные требования к методике оценки киберустойчивости в зависимости от класса IoT-системы и уровня её критичности. Научная новизна работы заключается в формировании матрицы соответствия классов IoT-систем и методов оценки киберустойчивости с обоснованием выбора по критериям отраслевого домена, уровня критичности и ключевого атрибута защищённости. Перспективным направлением дальнейших исследований представляется разработка унифицированного динамического индекса киберустойчивости с настраиваемыми весовыми коэффициентами компонентов, обеспечивающего количественное сопоставление систем различных классов в рамках единой методологии.

Библиографический список:

1. Ross R. et al. Developing cyber resilient systems: a systems security engineering approach. – National Institute of Standards and Technology, 2021. – NIST Special Publication (SP) 800-160 Vol. 2 Rev. 1.
2. International Electrotechnical Commission. IEC 62443-3-3:2013. Industrial communication networks – Network and system security – Part 3-3: System security requirements and security levels. – Geneva: IEC, 2013.
3. Alexander O., Belisle M., Steele J. MITRE ATT&CK for industrial control systems: Design and philosophy //The MITRE Corporation: Bedford, MA, USA. – 2020. – Т. 29. – С. 21-85.
4. The Open Group. Risk Analysis (O-RA) Standard. Version 2.0.1. – The Open Group, 2021.
5. Howard M., Lipner S. The security development lifecycle. – Redmond : Microsoft Press, 2006. – Т. 8.
6. UcedaVelez T., Morana M. M. Risk Centric Threat Modeling: process for attack simulation and threat analysis. – John Wiley & Sons, 2015.
7. Bian J. et al. Machine learning in real-time Internet of Things (IoT) systems: A survey //IEEE Internet of Things Journal. – 2022. – Т. 9. – №. 11. – С. 8364-8386.
8. Enisa E. Baseline security recommendations for IoT in the context of critical information infrastructures //European Union Agency for Cybersecurity Heraklion, Greece. – 2017.
9. Fagan M. et al. Foundational cybersecurity activities for IoT device manufacturers. – Gaithersburg, MD, USA : US Department of Commerce, National Institute of Standards and Technology, 2020.

Петренко С.А., Балябин А.А.
НТУ «Сириус», ФТ «Сириус»

МЕТОДИКА ОБЕСПЕЧЕНИЯ КИБЕРУСТОЙЧИВОСТИ ПЛАТФОРМ УПРАВЛЕНИЯ КИБЕРФИЗИЧЕСКИМИ ОБЪЕКТАМИ КИИ РФ

Введение. Критический анализ известных подходов обеспечения киберустойчивости киберфизических сетей и систем свидетельствует о недостаточности технологий (моделей, методов и средств) семантического контроля функционирования облачных платформ (ОП) управления киберфизическими системами (КФС) в условиях беспрецедентного роста угроз информационной безопасности. Поэтому, потребовалась разработка новой методики обеспечения киберустойчивости упомянутых объектов

исследования на основе *теории подобия*, которая позволяет учитывать семантические аспекты облачных машинных вычислений [1,2].

Общая идея предлагаемой методики заключается том, что исходную вычислительную программу ОП управления КФС становится возможным преобразовать в программу с кибериммунитетом, и на ее основе - осуществить синтез эталонной семантической модели. Здесь необходимо оценивать текущие значения вероятности достижения цели функционирования $P_{\text{цели}i}$ и времени выполнения программного цикла $T_{\text{вып}i}$, а также осуществлять контроль семантической корректности для подмножества линейных блоков упомянутой вычислительной программы. Для этого предлагается воспользоваться так называемым коэффициентом покрытия кибериммунитета $k_{\text{покр}i}$, который и позволит обеспечить выполнение требований $P_{\text{цели}i} \geq P^{\text{тр}}_{\text{цели}i}$ и $T_{\text{вып}i} \leq T^{\text{тр}}_{\text{вып}i}$ в условиях кибератак злоумышленников. Здесь при обнаружении кибератак злоумышленников необходимо будет осуществить возврат программы облачных машинных вычислений в начальное состояние и формировать так называемую «*память кибериммунитета*» [1,2].

Основные этапы предлагаемой методики. К упомянутым этапам предлагаемой методики относятся.

Этап 1. Преобразование исходной программы ОП управления КФС в программу с кибериммунитетом.

Целью данного этапа является преобразование исходной программы ОП управления КФС в программу с кибериммунитетом, построение ее эталонной семантической модели и формирование перечня линейных блоков, ранжированного по степени критичности нарушений. Описание входных и выходных данных этапа представлено в таблице 1.

Таблица 1 – Описание входных и выходных данных 1-го этапа

Входные данные	Вычислительная программа ОП управления КФС, представимая управляющим графом $\Gamma(B, D)$
Выходные данные	– Эталонная семантическая модель программы $M(S, A)$; – Ранжированный по степени критичности нарушений перечень линейных блоков B'.

Шаг 1.1. Исходная программа ОП управления КФС представляется в виде управляющего графа $\Gamma(B,D)$, где $B = \{b_j \mid j \in [1,k_i]\}$ – множество линейных блоков, $D = \{B \times B\}$ – множество связей по управлению. Каждый линейный блок b_j представляется последовательностью m_i арифметических выражений $b_j = \{b_{j1},...,b_{jm_i}\}$. Каждое арифметическое выражение в свою очередь имеет n_i операндов и представляется деревом разбора, которое в функциональном виде записывается как $b_{jl} = f_{jl1}(x_{jl1}, f_{jl2}(x_{jl2},...f_{jl,n_i-1}(x_{jl,n_i-1}, x_{jl,n_i})...))$

Шаг 1.2. Для каждого линейного блока b_j выполняется синтез матрицы инвариантов размерностей s_j. Совокупность матриц инвариантов, описывающих семантические состояния программы в линейных блоках, и переходов между ними составляет эталонную

Шаг 1.3. В линейные блоки управляющего графа программы b_j вносятся элементы избыточности в виде меток для последующего контроля семантики вычислений –

$b_j = (KT_j^{in}, b_{j1}, ..., b_{jm_i}, KT_j^{out})$, где KT_j^{in} и KT_j^{out} – метки начала и конца линейного блока соответственно.

Шаг 1.4. Множество линейных блоков программы B ранжируется по убыванию критичности нарушений таким образом, что $B' = \{b_1, ..., b_{j-1}, b_j, ..., b_{k_i}\}$ и критичность

на

Этап 2. Обнаружение нарушений семантики облачных машинных вычислений

Целью данного этапа является контроль нарушений семантики вычислительных структур во время выполнения программы ОП управления КФС с кибериммунитетом в условиях кибератак злоумышленников. Описание входных и выходных данных этапа представлено в таблице 2.

Таблица 2 – Описание входных и выходных данных 2-го этапа

Входные данные	– Требования к показателям устойчивости $P_{\text{цели}i}^{\text{тр}}$ и $T_{\text{вып}i}^{\text{тр}}$; – Эталонная семантическая модель программы $M(S, A)$; – Ранжированный по степени критичности нарушений перечень линейных блоков B'.
Выходные данные	Вывод о наличии нарушения семантики вычислительных структур и необходимости восстановления штатного функционирования

в

b_{j-1}

выше;

Шаг 2.1. Выполняется оценка текущих значений показателей вероятности достижения цели функционирования $P_{\text{цели}i}$ и времени выполнения программного цикла $T_{\text{вып}i}$. При $P_{\text{цели}i} < P_{\text{цели}i}^{\text{тр}}$ или $T_{\text{вып}i} > T_{\text{вып}i}^{\text{тр}}$ осуществляется синтез нового значения

ко

Шаг 2.2. Выполняется синтез необходимого $k_{\text{покр}i}^{\text{необх}}$ и достаточного $k_{\text{покр}i}^{\text{дост}}$ значений

эффициента покрытия кибериммунитета. Существование решения проверяется в

соответствии с введенным ранее критерием. При выполнении условия критерия

f_j

и

Шаг 2.3. На основе ранжированного перечня линейных блоков B' определяется

ще

Шаг 2.4. Выполняется синтез инварианта подобия s_j в текущем линейном блоке.

Полученный инвариант подобия проверяется на предмет соответствия эталонной семантической модели программы $M(S, A)$. При обнаружении нарушения управление

передается восстановителю штатного функционирования.

Этап 3. Восстановление штатного функционирования ОП управления КФС

Целью данного этапа является восстановление штатного функционирования вычислительной программы ОП управления КФС и сохранение сценариев противодействия кибератакам злоумышленников в памяти кибериммунитета. Описание входных и выходных данных этапа представлено в таблице 3.

Таблица 3 – Описание входных и выходных данных 3-го этапа

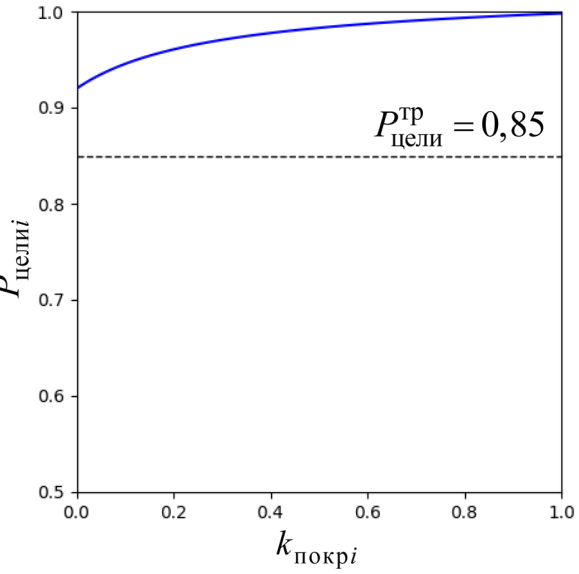
Входные данные	– Текущее семантическое состояние программы s_j'; – Вредоносные входные данные, обработка которых привела к возникновению нарушения $x' \in X_i^-$.

Выходные данные	Корректное семантическое состояние программы s_0

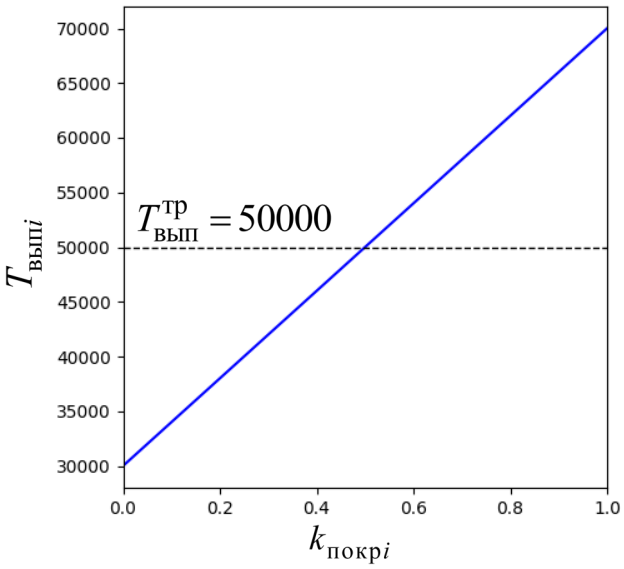
Шаг 3.1. Для текущего семантического состояния программы $s'_j \in M(S,A)$, осуществляется синтез восстанавливающего воздействия a_j, приводящего программу в корректное семантическое состояние $s'_j \xrightarrow{a_j} s_0$.

Шаг 3.2. Вредоносные входные данные $x' \in X_i^-$, обработка которых привела к возникновению нарушения семантики, сохраняются в памяти кибериммунитета и более не допускаются к обработке.

Оценка результативности. Проведем экспериментальное исследование устойчивости ОП управления КФС в условиях роста угроз безопасности информации.



а



б

На рисунке 1 представлены результаты исследования зависимости вероятности достижения цели функционирования ОП управления КФСна i-ом уровне $P_{\text{цели}i}$ и времени выполнения программного цикла $T_{\text{вып}i}$ от коэффициента покрытия кибериммунитета $k_{\text{покр}i}$, а на рисунке 2 – зависимости общей вероятности достижения цели $s_0 \in M(S,A)$,

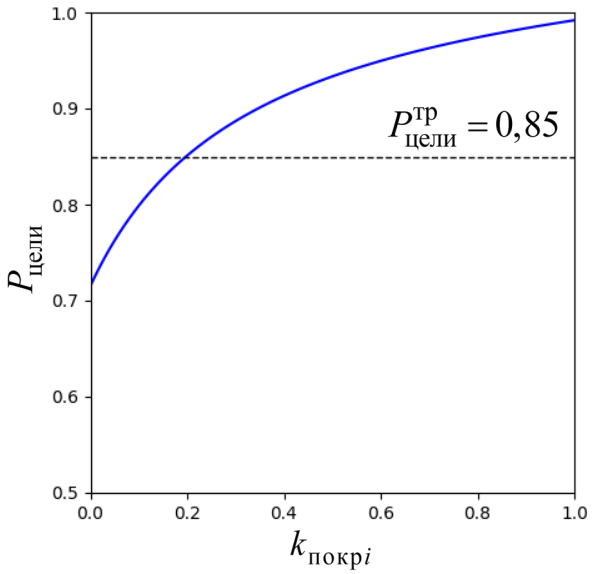


Рисунок 2 – Зависимость общей вероятности достижения цели функционирования ОП управления КФС $P_{\text{цели}}$ от $k_{\text{покр}i}$

На этих графиках представлено необходимые и достаточные значения коэффициента покрытия кибериммунитета: $k_{\text{покр}i}^{\text{необх}} = 0{,}1978$ и $k_{\text{покр}i}^{\text{дост}} = 0{,}5000$, удовлетворяющие введенному критерию, что свидетельствует о существовании решения.

Заключение. Научная новизна и практическая значимость методики заключается в проактивном характере противодействия кибератакам злоумышленников, нацеленном на своевременное обнаружение компьютерных атак злоумышленников и восстановления штатного функционирования облачных платформ (ОП) управления киберфизическими системами (КФС).

Исследование выполнено за счет гранта Российского научного фонда № 25-11-20037, https://rscf.ru/project/25-11-20037/

Список литературы:

1. Петренко С. А. Киберустойчивость индустрии 4.0 (научная монография при поддержке РФФИ)/ С. А. Петренко. – Санк-Петербург: Издательский Дом "Афина", 2020. – 256 с.
2. Петренко, С. А. Кибериммунология (научная монография при поддержке РФФИ / С. А. Петренко. – Санкт-Петербург : Афина, 2021. – 239 с.
3. Балябин А.А., Новиков В.А., Петренко С.А. Метод иммунного ответа на ранее неизвестные вредоносные воздействия // Методы и технические средства обеспечения безопасности информации. 2022. № 31. С. 11-13.
4. Балябин А.А. Метод защиты облачных платформ критической информационной инфраструктуры на основе кибериммунитета // Методы и технические средства обеспечения безопасности информации. 2025. № 34. С. 97-100.

Погорелов Н.С., Крундышев В.М.
Санкт-Петербургский политехнический университет Петра Великого, Санкт-Петербург

ИМИТАЦИОННАЯ МОДЕЛЬ РАСПРЕДЕЛЕННОГО РЕЕСТРА В ИНТЕЛЛЕКТУАЛЬНОЙ ТРАНСПОРТНОЙ СЕТИ УМНОГО ГОРОДА

Исследование выполнено за счет гранта Российского научного фонда №24-11-20005, https://rscf.ru/project/24-11-20005/; грант Санкт-Петербургского научного фонда (договор №24-11-20005 о предоставлении регионального гранта)

Технологии распределенного реестра являются одним из драйверов цифровизации мегаполисов. Основными преимуществами распределенных реестров являются прозрачность обработки данных, устойчивость к модификации информации и возможность организации доверенного взаимодействия между участниками без центрального управляющего узла. Однако, для подсистем умного города, построенных на базе распределенных реестров, актуален ряд атак, связанных с перегрузкой пула транзакций, увеличением задержек подтверждения транзакций, нарушением распространения данных между узлами сети и попытками недобросовестного поведения со стороны некоторых участников [1]. Для защиты от данных атак научные группы и команды разработчиков проводят исследования по созданию защитных механизмов, стойких протоколов и алгоритмов консенсуса [2, 3]. На этапе проверки гипотез (УГТ 3-5) для оценки эффективности механизмов защиты, как правило, используются средства моделирования, позволяющие воспроизводить различные сценарии функционирования систем и исследовать их поведение в контролируемых условиях. Однако, на сегодняшний день большая часть общедоступных имитационных моделей распределенных реестров разработана для традиционных сетей. В связи с этим возникает потребность создании модели, учитывающей особенности подсистем умного города: динамическую топологию, мобильность узлов, беспроводное соединение и ограниченные вычислительные ресурсы у конечных узлов сети. Целью данного исследования является разработка имитационной модели подсистемы умного города для оценки влияния атак на производительность распределенного реестра, анализа распространения транзакций и исследования изменения ключевых метрик системы.

Для создания имитационной модели распределенного реестра умного города выбран эмулятор Bitcoin-Simulator, предназначенный для моделирования Bitcoin-подобных сетей и позволяющий воспроизводить процессы распространения транзакций, формирования блоков и взаимодействия майнеров в распределенной среде. Благодаря гибкой настройке параметров сети, вычислительных мощностей узлов и задержек передачи сообщений, эмулятор предоставляет возможность исследовать поведение распределенного реестра в условиях низкой централизации и ограниченного хешрейта, характерных для «молодых» блокчейнов. Несмотря на то, что эмулятор Bitcoin-Simulator не ориентирован на масштабные сетевые эксперименты, он предоставляет удобную среду для моделирования атак и оценки устойчивости сети к нарушениям консенсуса [4]. Эмулятор позволяет реализовывать сценарии selfish-mining, перегрузки сети транзакциями, также он предоставляет возможность создавать собственные сценарии на основе базовых примеров.

С использованием эмулятора Bitcoin-Simulator была разработана модель распределенного реестра в интеллектуальной транспортной сети умного города, для визуализации был подключен модуль NetAnim. Разработанная модель позволяет воспроизвести сценарий использования распределенного реестра для оплаты проезда в общественном транспорте. На рис. 1 представлен фрагмент симуляции функционирования распределенного реестра в умном городе, где B – автобусы, M – злоумышленники (майнеры) и T – терминалы. Автобус генерирует платеж с идентификатором, временем и стоимостью проезда, после чего он отправляется на ближайший терминал. Терминал разбирает платеж: проверяет корректность полей, фильтрует дубликаты и помещает платеж

в очередь. Далее платеж пересылается назначенному ingress-майнеру. Злоумышленник добавляет транзакцию во внешний mempool и распространяет её по p2p-сети между другими майнерами. Майнер не начинает считать блок, пока не получит хотя бы одну транспортную транзакцию. После нахождения блока злоумышленник подтверждает транзакцию, блок распространяется по сети. Был реализован журнал событий: фиксируются создание платежа, приём и проверка терминалом, пересылка, получение майнером, распространение между майнерами и момент включения в блок. Дополнительно был разработан модуль, предназначенный для визуализации построения цепочки транзакций.

Рис. 1. Симуляция распределенного реестра в интеллектуальной транспортной сети

Таким образом, разработанная имитационная модель позволила воссоздать информационную инфраструктуру интеллектуальной транспортной сети, включающую транспортные узлы, терминалы обработки платежей, узлы-майнеры и механизм распространения транзакций. Разработанная модель позволяет воспроизвести атаки типа flooding attack и selfish-mining attack, а также оценить эффективность механизмов защиты на основе ограничения нагрузки, контроля распространения транзакций и управления обработкой платежей с последующим сравнением метрик нормального, атакуемого и защищённого состояний сети.

Список литературы:

1. Петренко, С. А. Модель квантовых угроз безопасности информации для национальных блокчейн-экосистем и платформ / С. А. Петренко, А. А. Балябин // Вопросы кибербезопасности. – 2025. – № 1(65). – С. 7-17.
2. Коноплев, А. С. Защита блокчейн-систем умного города от атаки эгоистичного майнинга / А. С. Коноплев, М. О. Калинин // Проблемы информационной безопасности. Компьютерные системы. – 2025. – № 2(64). – С. 60-70.
3. Балябин, А. А. Метод обеспечения киберустойчивости блокчейн-платформ на основе кибериммунитета / А. А. Балябин, С. А. Петренко // Вопросы кибербезопасности. – 2025. – № 6(70). – С. 127-139.
4. Gervais A. et al. On the security and performance of proof of work blockchains //Proceedings of the 2016 ACM SIGSAC conference on computer and communications security. – 2016. – С. 3-16.

Правиков Д.И. Цурцумия А.О.
РГУ нефти и газа (НИУ) имени И.М. Губкина, Москва

ПРИМЕНЕНИЕ КОНСТРУКТИВНОЙ БЕЗОПАСНОСТИ ДЛЯ ПОВЫШЕНИЯ БЕЗОПАСНОСТИ УПРАВЛЕНИЯ НЕФТЕПЕРЕКАЧИВАЮЩЕЙ СТАНЦИЕЙ

В настоящее время для защиты от компьютерных атак на системы промышленной автоматизации используют наложенные средства защиты информации, применение которых требует, как показывает практика, решения целого ряда технических задач, Вместе с тем, АО «Лаборатория Касперского» предлагает в определенной степени альтернативный подход, связанный с разработкой систем управления, обладающих свойствами конструктивной безопасности, описанный в [1]. Рассмотрим возможность применения указанного подхода к обеспечению безопасности объектов ТЭК на примере системы управления нефтеперекачивающей станции (НПС).

При анализе тактик и техник атак для систем промышленной автоматизации, описанных в матрице MITRE [7], можно сделать вывод, что, в общем случае, одним из более вероятных сценариев атаки является целенаправленное искажение данных, передаваемых между элементами системы управления технологическим процессом (исполнительными механизмами полевого уровня, ПЛК и SCADA системой). Как показано в [6], указанное воздействие может привести к существенным изменениям для управляемого технологического процесса.

Для реализации конструктивного подхода предлагается дополнить типовую схему АСУ ТП конструктивными элементами, нивелирующими возможное негативное воздействие на передаваемые данные [4]. На рисунке 1 представлена схема, которая включает в себя: элементы системы управления НПС, встроенные механизмы обеспечения достоверности и подлинности телеметрии, локальный мониторинг всей системы [2].

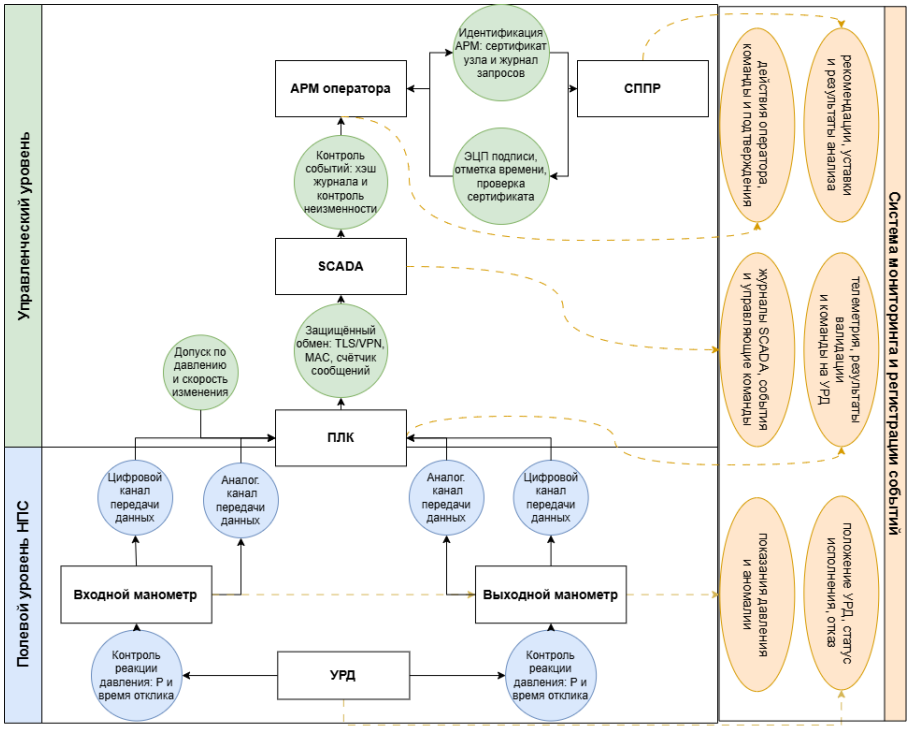


Рис. 1 – Архитектура передачи данных, защитных механизмов и централизованного мониторинга в системе управления НПС

Для связи манометров с ПЛК предложены дублирующие каналы передачи данных и механизм сравнения давления в трубе – это позволяет определять аномальные или

подменённые значения и не допускать дальнейшего их распространения. Для связи ПЛК и SCADA предусмотрен защищённый канал обмена данными, который включает в себя контроль целостности, аутентификацию канала и контроль последовательности сообщений. Для связи SCADA с АРМ введён механизм контроля критических событий и неизменности журналов – это позволяет предотвращать недостоверное отображение состояния станции для оператора. Для связи АРМ с СППР предложены механизмы защиты, которые включают в себя идентификацию источников запросов и журналирование обращений. Для защиты передачи данных от СППР до АРМ предусмотрены архитектурные решения в виде электронной подписи и отметки времени. Всё это позволяет подтвердить источник и актуальность данных и передаваемые задания [3].

Отдельным ключевым элементом в данной архитектуре является система мониторинга и регистрации событий. Она отвечает за целостную наблюдаемость системы, получая данные от всех элементов системы, и обеспечивает защиту отдельных связей между ними [5].

Предложенные решения напрямую соотносятся с положением ГОСТ Р 72118–2025, в частности с шаблонами «Монитор», «Иерархия доверия» и «Выделенный обработчик для очистки данных» [1]. Данные шаблоны реализовывают и обосновывают, в рамках подхода конструктивной безопасности, централизованный мониторинг, поуровневую проверку достоверности информации и фильтрацию нетипичных сигналов.

Таким образом, для объектов ТЭК возможно использовать конструктивную безопасность как подход, меняющий свойства архитектуры системы управления, где механизмы контроля, проверки достоверности телеметрии, подтверждение подлинности информации и мониторинг всей системы встроены в сами процессы передачи данных. Это существенно снижает риск передачи искажённой информации в следствие чего, предотвращается формирование некорректных управленческих решений, а также обеспечивается переход к целостному контролю функционирования НПС.

Библиографический список:

ОСТ Р 72118—2025. Защита информации. Системы с конструктивной информационной безопасностью. Методология разработки. — Москва: Стандартинформ, 2025.

еккер Л. М., Назаренко А. В. Нефтеперекачивающая станция бесперебойной работы. — 2016.

елых В. А. Модернизация автоматизированной системы управления нефтеперекачивающей станции. — 2023.

елоусов А. Н. Автоматизация нефтеперекачивающей станции. — 2017.

оростелев Д. А., Левая М. Н., Левый И. С. Автоматизация процесса мониторинга и регулирования давления в нефтепроводе // Вестник Астраханского государственного технического университета. Серия: Управление, вычислительная техника и информатика. — 2017. — № 1. — С. 37–47.

евцов В. Ю., Правиков Д. И. Применение сетей Петри при моделировании атак на системы АСУ ТП // Вестник современных цифровых технологий. — 2022. — № 13. — С. 32–37. — EDN RYUSGM.

7. Xiong W. [et al.]. Cyber security threat modeling based on the MITRE Enterprise ATT&CK Matrix // Software and Systems Modeling. — 2022. — Vol. 21, no. 1. — P. 157–177.

Фетеску Н.Г.
РГУ нефт и и газа (НИУ) имени И.М. Губкина, Москва

ПРОГНОЗИРОВАНИЕ ИЗМЕНЕНИЯ УРОВНЯ ЗАЩИЩЕННОСТИ СЕГМЕНТА АСУ ТП НА ОСНОВЕ ПОТОКА ВЫЯВЛЯЕМЫХ УЯЗВИМОСТЕЙ

Объекты критической информационной инфраструктуры, функционирующие в составе автоматизированных систем управления технологическими процессами предприятий топливно-энергетического комплекса, требуют не только первичной оценки защищенности, но и последующего прогноза ее изменения. Разовая оценка состояния объекта до или после внедрения средств защиты информации не отражает динамику уязвимого фона, поскольку новые уязвимости продолжают выявляться даже при неизменной архитектуре системы.

Целью работы является формирование подхода к прогнозированию изменения уровня защищенности сегмента АСУ ТП на основе потока выявляемых уязвимостей. Методический замысел опирается на работы, посвященные количественной оценке информационной безопасности и показателям состояния защищенности объектов нефтегазовой отрасли [1, 2]. В качестве источника динамических данных может использоваться Банк данных угроз безопасности информации ФСТЭК России.

Для модельного эксперимента рассматривается типовой сегмент промышленной автоматизации, включающий серверы верхнего уровня, автоматизированные рабочие места, инженерные станции, сетевое оборудование, средства удаленного доступа и средства защиты информации. Каждому компоненту инфраструктуры сопоставляется множество релевантных уязвимостей, выявленных за период наблюдения T.

Интегральный показатель накопленной уязвимости сегмента в момент времени t предлагается определять следующим образом:

$$U(t) = \sum_i w_i \cdot e_i \cdot \sum_{j \in V_i(t)} q_{ij} \cdot (1 - c_{ij})$$

где:

w_i – вес i-го компонента инфраструктуры,

e_i – коэффициент экспозиции компонента,

$V_i(t)$ – множество уязвимостей, релевантных компоненту к моменту времени t,

q_{ij} – нормированная критичность j-й уязвимости,

c_{ij} – степень компенсации уязвимости действующими мерами защиты.

Уровень защищенности сегмента может быть представлен как нормированная величина, обратная накопленной уязвимости:

$$Z(t) = \max\left\{0, 1 - \frac{U(t)}{U_{доп}}\right\}$$

где $U_{доп}$ – предельно допустимый уровень накопленной уязвимости.

Тогда защищенность объекта до внедрения СЗИ описывается значением $Z_0 = Z(t_0)$, защищенность после внедрения – $Z_1 = Z(t_1)$, а прогнозное состояние через заданный период – $Z(t_1 + T)$. При неизменном составе защитных мер значение c_{ij} остается постоянным, а рост множества $V_i(t)$ приводит к увеличению $U(t)$ и снижению $Z(t)$.

Для оценки скорости снижения защищенности могут использоваться показатели интенсивности поступления релевантных уязвимостей и среднего темпа деградации защищенности:

$$\lambda_v = \frac{N_{рел}(T)}{T};\ \gamma_z = \frac{Z(t_1) - Z(t_1 + T)}{T}$$

где $N_{рел}(T)$ – число уязвимостей, применимых к компонентам сегмента за период T, λ_v – интенсивность поступления релевантных уязвимостей, γ_z – средний темп снижения защищенности. Дополнительно может определяться расчетный срок $T_{кр}$, при котором $Z(t)$ достигает минимально допустимого уровня Z_{min}.

Предлагаемый подход позволяет получить динамические показатели, необходимые для последующей методологии внедрения СЗИ в АСУ ТП предприятий ТЭК: уровень защищенности до внедрения защитных мер, уровень защищенности после внедрения, прогнозное значение защищенности через заданный интервал времени, интенсивность появления релевантных уязвимостей и коэффициент устаревания защитных мер. Практическая значимость заключается в возможности обосновывать сроки повторной оценки защищенности, актуализации модели угроз и корректировки состава внедренных СЗИ.

Список литературы:

1. Правиков Д.И., Мурашкин В.А. Подходы к количественной оценке информационной безопасности на предприятии ТЭК // Проблемы управления безопасностью сложных систем: материалы XXXII междунар. конф. М.: ИПУ РАН, 2024. С. 212–217. URL: https://elibrary.ru/item.asp?id=79446201.
2. Правиков Д.И., Мурашкин В.А. Показатель состояния защищенности на объектах нефтегазовой отрасли // Кибернетика и информационная безопасность «КИБ-2024»: сборник науч. трудов Второй Всерос. науч.-техн. конф. М.: НИЯУ МИФИ, 2024. С. 26–27. URL: https://elibrary.ru/item.asp?id=75062623.
3. Сухих Я.А., Правиков Д.И., Кузичкин А.А. Разработка защищенных архитектур автоматизированных систем управления технологическими процессами // Безопасность информационных технологий. 2020. Т. 27. № 2. С. 97–117. DOI: 10.26583/bit.2020.2.08.

3. СОВРЕМЕННЫЕ ТЕХНОЛОГИИ КИБЕРБЕЗОПАСНОСТИ

Агафонов А.С.
Уральский государственный юридический университет имени В. Ф. Яковлева, Екатеринбург

ЗАРУБЕЖНЫЙ ОПЫТ РЕГУЛИРОВАНИЯ В ОБЛАСТИ ОБЕСПЕЧЕНИЯ ИНФОРМАЦИОННОЙ БЕЗОПАСНОСТИ И ЦИФРОВОГО СУВЕРЕНИТЕТА ГОСУДАРСТВА

В настоящем докладе используется сравнительная методология для изучения опыта регулирования в области обеспечения информационной безопасности и цифрового суверенитета в различных государствах[1]. Выбор стран для сравнительного исследования представляется достаточным, чтобы формулировать научные выводы универсального характера. Особенно если учитывать ранее обозначенный Президентом России факт, в соответствии с которым на сегодняшний день цифрового суверенитета достигли лишь США, Китай и Россия. Более того, география исследования охватывает иные современные государства, которые расположены в разных частях света, принадлежат различным правовым семьям, демонстрируют различные культурный, технологический и исторический контексты. Сравнительная методология не только создаёт условия для расширения кругозора за счёт изучения конструкций и практик, не представленных в родной юрисдикции, но и даёт возможность изучить проблемы, которые могли бы возникнуть, если бы была избрана иная стратагема построения того или иного института.

Европейский Союз (на примере Франции)

В контексте стран Европейского Союза (далее – ЕС) намечается концепция, при которой информационная безопасность и цифровой суверенитет рассматриваются не в части отдельных государств, а в контексте самого ЕС. Так, во Франции создан специальный Институт цифрового суверенитета. Он направлен на: обобщение и систематизацию информации, связанной с управлением Интернета; привлечение граждан и представителей экспертного сообщества к вопросам цифрового суверенитета; формирование предложений по мерам технологического, правового и политического характера в части обеспечения цифрового суверенитета; создание синергии по вопросам цифрового суверенитета на территории ЕС. Французский опыт обеспечения информационной безопасности и цифрового суверенитета позволяет выделить и специализированный государственный орган – *Agence nationale de la sécurité des systèmes d'information* (далее – *ANSSI*), который осуществляет координацию с министерствами и другими институциональными субъектами, участвующими в государственной политике по обеспечению информационной безопасности. В части регулирования отметим, что во Франции (как и, к примеру, в Германии) применяется подход, используемый на уровне ЕС, из-за чего национальное

[1] В настоящем докладе представлена лишь часть результатов научных изысканий, касающихся зарубежного опыта регулирования в области обеспечения информационной безопасности и цифрового суверенитета. Аккумулируются некоторые тезисы и делается умозаключение. Краткий стиль изложения продиктован требованиями к объёму доклада, который без учёта этого примечания составляет 2 страницы. Подробные результаты и опыт других стран (Северной Кореи, Австралии и т.д.), могут быть представлен в рамках 35-ой конференции МиТСОБИ.

законодательство приводится в соответствие с нормами права, которые имеют место на уровне ЕС.

Таким образом, можно сделать вывод о том, что страны ЕС идут по концепции построения цифрового суверенитета ЕС в целом, а не цифрового суверенитета отдельных европейских государств, что ярко выражено в представленном опыте Франции.

Индия

Меры, принимаемые для обеспечения информационной безопасности и цифрового суверенитета Индии, характеризуются превалированием императивного начала. Регулирование в Индии строится на основе программы *Digital India*, которая направлена на расширение цифровых прав граждан в части предоставления государственных услуг и развитии цифровой инфраструктуры страны. Отметим, что в Индии ведётся работа по созданию Центра передового опыта в области искусственного интеллекта и открытию около 50 тыс. лабораторий *Atal Tinkering Labs* в государственных школах, что следует из бюджета на 2025–2026 финансовый год.

Таким образом, опыт Индии позволяет изучить, как в рамках одной концепции цифрового суверенитета происходит поиск баланса публичных и частных интересов в части соотношения императивного и диспозитивного. Концепция Индии основывается на построении собственного цифрового суверенитета, а не на его совместном создании с другими государствами как, к примеру, в ЕС.

Кит ай

Китай нацелен на: создание открытого и безопасного национального рынка цифровых продуктов и укрепление использования информационных ресурсов; ускорение внедрения искусственного интеллекта (в части программы «искусственный интеллект+» (кит.: «*人工智能+*») и его интеграцию с промышленным развитием, культурным строительством. Основой для таких рекомендаций выступает *数字中国建设整体布局规划* (далее – Общий план строительства цифрового Китая), где построение цифрового Китая рассматривается как создание нового национального конкурентного преимущества. Общий план строительства цифрового Китая предусматривает глубокую интеграцию цифровых технологий по правилу «пять в одном (кит.: «*五位一体*»). В Китае сформирована собственная сетевая инфраструктура, функционирующая с помощью национальных цифровых технологий, и, позволяющая обеспечить функционировании Китая в цифровом пространстве даже в случае его отключения от международной сети «Интернет», из чего следует умозаключение об обладании Китаем цифрового суверенитета.

Таким образом, опыт Китая позволяет изучить, как в рамках концепции цифрового суверенитета государство стремится к независимому существованию в национальном и международном цифровом пространствах, при сохранении национальной идентичности и публично-правового контроля с учётом различных условий (политических, социальных, экономических и т.д.) собственного развития.

США

Цель США заключается в том, что именно американские технологии и стандарты двигали мир вперед. США стремится защитить собственный государственный суверенитет и не допустить его «размывания» транснациональными компаниями и международными организациями. Отметим документ, который связан с регулированием искусственного интеллекта в США – *AI Action Plan* (План действий США в области искусственного интеллекта) .

Таким образом, опыт США позволяет изучить, как в рамках концепции цифрового суверенитета государство стремится сохранить своё доминирующее положение в цифровом пространстве. США нацелено сформировать цифровое пространство, которое будет полностью соответствовать американским ценностям, и, место в котором будет только для тех субъектов права, которые будут разделять эти ценности.

Выводы

Элементы изученных концепций, с помощью которых строится цифровой суверенитет обозначенных государств, могут быть и должны быть рассмотрены на предмет рецепции в доктрину информационной безопасности и концепцию цифрового суверенитета Российской Федерации постольку, поскольку они отражают разносторонние подходы к достижению полной цифровой самостоятельности, в том числе основанные на:

1) поиске баланса публичных и частных интересов в части соотношения императивного и диспозитивного в рамках концепции цифрового суверенитета экономически-развивающегося государства;

2) сохранении национальной идентичности и публично-правового контроля с учётом различных условий собственного развития;

3) формировании цифрового пространства, которое будет полностью соответствовать его ценностям, и, место в котором будет только для субъектов права, разделяющих эти ценности.

Аксёнов К.Д.
Санкт-Петербургский государственный университет телекоммуникаций им. проф. М. А. Бонч-Бруевича, Санкт-Петербург
ООО Пространство интеллектуальных решений, Новороссийск

АНАЛИЗ МОДЕЛЕЙ ОБЕСПЕЧЕНИЯ БЕЗОПАСНОСТИ МЕДИЦИНСКИХ ДАННЫХ НА ЭТАПАХ ИХ ЖИЗНЕННОГО ЦИКЛА

Современные медицинские информационные системы характеризуются постоянным ростом объёмов диагностических данных, широким распространением телемедицинских технологий и интеграцией медицинских организаций в объекты критической информационной инфраструктуры. В условиях цифровизации здравоохранения особую значимость приобретают вопросы обеспечения конфиденциальности, целостности и доступности медицинской информации на всех этапах жизненного цикла медицинских данных. При этом растущее многообразие применяемых средств защиты — от криптографических механизмов до методов стеганографии и цифровых водяных знаков — требует не только разработки новых способов защиты, но и формализованных подходов к оценке их эффективности и соответствия установленным нормативным требованиям.

Анализ исследований показывает, что существующие подходы к обеспечению безопасности медицинской информации преимущественно ориентированы либо на отдельные этапы обработки данных, либо на отдельные методы защиты, при этом вопросы количественной оценки защищённости проработаны фрагментарно. Так, в работе Rindfleisch были систематизированы базовые угрозы конфиденциальности медицинской информации в условиях растущей информатизации здравоохранения и обозначен круг требований к организационно-техническим мерам её защиты, однако вопросы количественной оценки эффективности таких мер не рассматривались [1]. В работе Khaloufi и соавт. систематизированы угрозы и механизмы защиты на этапах жизненного цикла больших данных в здравоохранении, однако предложенная модель носит описательный характер и не содержит измеримых критериев оценки эффективности защитных мер [2]. В исследованиях Tan и соавт. для оценки методов цифрового водяного знака в DICOM-изображении применяются метрики PSNR и BER, что ограничивает оценку параметрами одной операции встраивания и не связывает её с требованиями к информационной системе в целом [3]. В обзоре Eichelberg и соавт. угрозы безопасности PACS-систем каталогизированы, однако не приведены к измеримым показателям защищённости инфраструктуры медицинской визуализации [4]. В монографии Huang

подробно описаны архитектура и процессы PACS-систем и обработки медицинских изображений, однако вопросы оценки защищённости таких систем затрагиваются преимущественно на уровне общих принципов организации доступа и резервирования данных [5]. В работе Weiskopf и Weng предложены подходы к оценке качества данных электронных медицинских записей по группам показателей полноты, корректности и согласованности, которые ориентированы на задачи повторного использования данных и не учитывают требований к защищённости информации [6].

Согласно Meingast и соавторам, обеспечение безопасности медицинских информационных технологий требует комплексного подхода, включающего управление доступом и контроль целостности на различных этапах обработки информации, тогда как формализация критериев такой оценки остаётся открытой задачей [7]. В работах Красова и соавт. отмечается, что медицинские информационные системы офтальмологического профиля являются объектами критической информационной инфраструктуры и требуют реализации специализированных механизмов защиты в соответствии с требованиями 187-ФЗ и нормативных документов ФСТЭК России [8]. Bose и Marijan указывают на возрастающий интерес к моделям защиты медицинских данных, охватывающим все процессы их жизненного цикла, в том числе долговременное хранение и архивирование информации, но подчёркивают отсутствие единой методики оценки таких моделей [9].

В качестве нормативной основы при формировании критериев оценки защитных мер на этапах жизненного цикла медицинских данных могут использоваться: ГОСТ Р 52636-2006, устанавливающий общие требования к ведению электронных персональных медицинских записей; ГОСТ Р ИСО 14721 (OAIS), определяющий эталонную модель долговременного архивного хранения цифровых объектов; ГОСТ Р ИСО/МЭК 12207, описывающий процессы жизненного цикла программных средств. Требования к защите задаются 187-ФЗ, 152-ФЗ и приказами ФСТЭК России № 235 и № 239, определяющими категории значимости объектов критической информационной инфраструктуры и соответствующие меры защиты информации. Совокупность этих документов формирует нормативный контур, в рамках которого должна выполняться оценка применения методов защиты медицинских данных.

Под жизненным циклом медицинских данных в настоящей работе рассматриваются процессы получения, обработки, передачи, оперативного хранения, архивирования, восстановления и обезличивания информации. В работе Khaloufi и соавт. жизненный цикл медицинских данных включает этапы Data Collection, Data Transformation, Data Modeling, Knowledge Creation и Data Storage, для каждого из которых характерны собственные угрозы информационной безопасности, такие как spoofing-атаки, phishing, атаки на целостность данных, re-identification threats, content-based attacks и data mining attacks, а также соответствующие механизмы защиты [2]. Для оценки защищённости на каждом из этапов используют собственные группы показателей: стойкость криптографических механизмов и характеристики систем разграничения доступа — на этапах передачи и хранения; метрики PSNR, SSIM, ёмкости скрытого канала (Encoding Capacity) и устойчивости к стегоанализу (Steganalysis Robustness) — для оценки стеганографических методов защиты медицинских изображений; модели сохранности и аутентичности из OAIS — для этапа архивирования. В исследовании Штеренберга подчёркивается необходимость интеграции механизмов информационной безопасности непосредственно в архитектуру интеллектуальных медицинских систем и учёта показателей киберустойчивости при оценке защищённых систем искусственного интеллекта [10]. Однако перечисленные группы метрик применяются изолированно: отсутствует формализованный переход от значений отдельных показателей к интегральному заключению о соответствии системы требованиям критической информационной инфраструктуры с учётом класса значимости объекта защиты.

Таким образом, перспективным направлением исследований является разработка модели оценки применения методов скрытого встраивания данных на этапах жизненного цикла медицинских данных. Такая модель должна обеспечивать формализованный переход от

значений метрик защищённости (PSNR, SSIM, Encoding Capacity, Steganalysis Robustness) к заключению о соответствии требованиям критической информационной инфраструктуры и законодательства о персональных данных с учётом класса значимости объекта защиты, учитывать процессы долговременного архивирования и восстановления информации, обеспечивать сохранение диагностической значимости медицинских изображений и допускать применение в задачах аттестации медицинских информационных систем.

Список литературы:

1. Rindfleisch T.C. Privacy, Information Technology, and Health Care // Communications of the ACM. 1997. Vol. 40(8). P. 92–100. DOI: 10.1145/257874.257896.
2. Khaloufi H., Abouelmehdi K., Beni-Hssane A., Saadi M. Security model for Big Healthcare Data Lifecycle // Procedia Computer Science. — 2018. — Vol. 134. — P. 270–277. — DOI: 10.1016/j.procs.2018.07.183.
3. Tan J., Ng S., Xu H. Security Protection of DICOM Medical Images Using Dual-Layer Reversible Watermarking with Tamper Detection Capability// Journal of Digital Imaging. 2011. Vol. 24(2). P. 311–321. DOI: 10.1007/s10278-010-9298-4.
4. Eichelberg M., Kleber K., Kämmerer M. Cybersecurity in PACS and Medical Imaging: An Overview // Journal of Imaging. 2020. Vol. 6(12). Article 138. DOI: 10.3390/jimaging6120138.
5. Huang H. K. PACS and Imaging Informatics: Basic Principles and Applications. — 2-е изд. — Hoboken : Wiley-Blackwell, 2010. — 1048 p. — ISBN 978-0470560518.
6. Weiskopf N.G., Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research // Journal of the American Medical Informatics Association. 2013. Vol. 20(1). P. 144–151. DOI: 10.1136/amiajnl-2011-000681.
7. Meingast M., Roosta T., Sastry S. Security and Privacy Issues with Health Care Information Technology // Proceedings of the 28th IEEE EMBS Annual International Conference. 2006. P. 5453–5458. DOI: 10.1109/IEMBS.2006.260060.
8. Красов А. В., Лансере Н. Н., Фадеев И. И., Гельфанд А. М., Лесневский М. В. Типовые офтальмологические информационные системы, являющиеся объектами критической информационной инфраструктуры // Офтальмохирургия. — 2022. — № 4S. — С. 85–91. — DOI: 10.25276/0235-4160-2022-4S-85-91.
9. Bose, S., Marijan, D. A survey on privacy of health data lifecycle: a taxonomy, review, and future directions. Int. J. Inf. Secur. 25, 33 (2026). https://doi.org/10.1007/s10207-025-01169-y.
10. Штеренберг С. И. Методика построения защищенных систем искусственного интеллекта для проведения электроретинографии в офтальмологии // Офтальмохирургия. — 2022. — № 4S. — С. 51–57. — DOI: 10.25276/0235-4160-2022-4S-51-57.

Антонова Г.В.

Санкт-Петербургский государственный морской технический университет, Санкт-Петербург

НЕДОСТАТКИ НОРМАТИВНО-ПРАВОВОГО РЕГУЛИРОВАНИЯ ИНФОРМАЦИОННОЙ БЕЗОПАСНОСТИ ПРИ ИСПОЛЬЗОВАНИИ СРЕДСТВ РАДИОЭЛЕКТРОННОГО ПОДАВЛЕНИЯ

В условиях роста угроз использования беспилотных летательных аппаратов (БПЛА), контрабандных мобильных устройств в исправительных учреждениях и

несанкционированного доступа к каналам связи, средства подавления (глушилки) становятся востребованным инструментом обеспечения безопасности. Однако их применение сопряжено с серьёзными правовыми коллизиями и техническими рисками, а нормативно-правовая база демонстрирует системные недостатки, требующие научного осмысления и практических рекомендаций по совершенствованию.

1. Проблема избыточной бюрократизации согласования и «вымывания» рабочих диапазонов.

Легальное применение средств подавления (глушилок) в Российской Федерации требует прохождения многоступенчатой процедуры согласования с 7–10 государственными органами, включая:

- ГКРЧ (Государственная комиссия по радиочастотам);
- Роскомнадзор;
- Минобороны РФ;
- ФСБ РФ;
- ФСО РФ;
- МВД РФ;
- Росавиацию;
- региональные администрации связи;
- в ряде случаев — ОАО «РЖД» и РТРС.

Каждая инстанция в рамках своих полномочий и требований по защите ведомственных и критических систем связи исключает из разрешённой полосы частот диапазоны, задействованные для нужд государственного управления, обороны и безопасности. Срок согласования составляет от 3 до 6 месяцев.

В результате итеративного «вырезания» частот итоговая разрешённая полоса работы глушилки существенно сужается, а её техническая эффективность падает. Возникает парадокс: средство, легально предназначенное для нейтрализации угроз (контрабандные телефоны в колониях, несанкционированный запуск БПЛА), после согласований утрачивает способность полноценно подавлять целевые сигналы именно в той среде, где реально работают нарушители.

Это приводит к формальному соблюдению закона («разрешение получено»), но фактической неспособности выполнить защитную функцию.

2. Необходимость нового подхода к регулированию.

Требуется разработка сбалансированного нормативно-правового механизма, который, с одной стороны, сохранит контроль государства над использованием спектра и защиту критически важных диапазонов, а с другой — не будет препятствовать корректной работе глушилки на заявленных для подавления частотах.

Возможные направления решения:

ведение типовых (упрощённых) разрешений для средств подавления в ограниченных пространствах (внутри зданий, на огороженной территории), где риск воздействия на внешние системы сведён к минимуму.

ереход от сплошного «вырезания» частот к нормированию уровня побочного излучения за пределами охраняемой зоны (энергетический критерий вместо запретительного).

оздание «белых списков» частот (разрешены все диапазоны, кроме явно перечисленных критических — ведомственные, военные, аварийно-спасательные), что снижает объём согласований.

ифровизация процедуры и введение «единого окна» с фиксированным сроком согласования, чтобы исключить многомесячные бюрократические цепочки.

3. Предлагаемые изменения в законодательство.

3.1. Федеральный закон № 126-ФЗ «О связи».

Статья для изменения: статья 24 «Использование радиочастотного спектра»

Предлагаемое изменение: дополнить пунктом о возможности получения типовых (упрощённых) разрешений для средств подавления в ограниченных пространствах (внутри зданий, на огороженной территории) с фиксированным перечнем допустимых диапазонов и мощностей.

3.2. Постановление Правительства РФ № 1002 «О государственной радиочастотной службе».

Предлагаемое изменение: внести поправку, предусматривающую переход от сплошного «вырезания» частот к нормированию уровня побочного излучения за пределами охраняемой зоны (энергетический критерий вместо запретительного).

3. Решения ГКРЧ (типовая форма).

Предлагаемое изменение: утвердить «белые списки» частот — разрешены все диапазоны, кроме явно перечисленных критических (ведомственные, военные, аварийно-спасательные), что снижает объём согласований.

3.4. Административный регламент Роскомнадзора.

Предлагаемое изменение: внедрить цифровизацию процедуры и принцип «единого окна» с фиксированным сроком согласования (не более 45 рабочих дней).

Таблица 1. Введение классификации средств подавления

Класс	Характеристики	Порядок
Класс А	Маломощные, для помещений	Уведомительный порядок
Класс Б	Для открытых территорий	Разрешительный порядок с энергетическими критериями

Заключение

Предлагаемый подход позволит:

- сохранить государственный контроль над радиочастотным спектром;
- не превращать легальную глушилку в декоративное устройство;
- существенно сократить сроки согласования (с 3–6 месяцев до 1–2 месяцев);
- обеспечить реальную эффективность средств подавления на защищаемых объектах;
- снизить риски «побочного поражения» гражданских систем связи.

Антропов Д.С., Марченко И.В.

Национальный исследовательский ядерный университет «МИФИ», Москва

МЕТОД ФОРМИРОВАНИЯ ОБЪЯСНЕНИЙ РЕЗУЛЬТАТОВ КЛАССИФИКАЦИИ ИНЦИДЕНТОВ ИНФОРМАЦИОННОЙ БЕЗОПАСНОСТИ

Рост объема событий безопасности и усложнение сценариев атак приводят к тому, что для аналитика SOC недостаточно только автоматически присвоенного класса инцидента. Практически значимым становится не только результат классификации, но и его интерпретация, так как именно она определяет доверие к системе, скорость верификации решения и обоснованность выбора ответных мер. Поэтому в интеллектуальных системах поддержки принятия решений в области информационной безопасности требуется формирование краткого, структурированного и воспроизводимого объяснения результата классификации.

Пусть инцидент информационной безопасности представлен упорядоченной последовательностью событий и набором извлеченных признаков. На этапе классификации для каждого класса угроз $k_j \epsilon K$ уже вычислена интегральная оценка принадлежности $S(j)$. Тогда задача объяснения состоит в формировании результата, который включает выбранный класс инцидента $k*$, ближайший конкурирующий класс $k(2)$, запас уверенности решения Δ, множества усиливающих и ослабляющих признаков, перечень решающих признаков относительно конкурирующего класса и причину передачи на экспертный анализ при недостаточной уверенности. Выбранный класс определяется стандартным правилом максимума:

$$k* = arg\ max\ S(j), k_j \epsilon K$$

Для оценки устойчивости решения дополнительно определяется второй по величине кандидат:

$$k(2) = arg\ max\ S(j), k_j \epsilon K, k_j \neq k*$$

Запас уверенности рассчитывается как:

$$\Delta = S(k*) - S(k(2))$$

Введение величины Δ позволяет отличать уверенные решения от пограничных случаев. Если Δ мало, то даже при совпадении итогового класса с максимумом интегральной оценки аналитик получает сигнал о наличии альтернативной интерпретации рассматриваемого инцидента.
Для каждого признака x_f, участвующего в вычислении частных оценок, известен вклад c_f^k в результат для класса k и штраф p_f^k, отражающий степень несоответствия ожидаемому профилю класса. Тогда усиливающие признаки определяются как признаки с максимальными значениями c_f^{k*}, а ослабляющие признаки - как признаки с максимальными значениями p_f^{k*}. Однако для аналитика важно понимать не только вклад признаков в абсолютном смысле, но и то, почему выбранный класс оказался предпочтительнее ближайшей альтернативы. Поэтому для каждого признака рассчитывается сравнительная разность вкладов:

$$\delta_f = c_f^{k*} - c_f^{k(2)}$$

После упорядочения по величине δ_f формируется множество решающих признаков D, то есть таких признаков, которые обеспечили преимущество выбранного класса над ближайшим конкурентом. Эта часть важна в ситуациях, когда инциденты имеют совпадающее техническое описание, но отличаются контекстом, временной структурой или совокупностью слабых признаков. Если итоговая оценка выбранного класса ниже заданного порога доверия θ, инцидент не считается окончательно классифицированным и передается на экспертный анализ. В этом случае к объяснению добавляется причина $r_a: S(k*) < \theta$, что позволяет использовать единый механизм как для объяснения автоматически принятых решений, так и для объяснения отказа от полностью автоматического выбора. Предлагаемый метод реализован в прототипе интеллектуальной системы поддержки принятия решений по инцидентам информационной безопасности. В качестве объяснения система формирует: выбранный класс, ближайший конкурирующий класс, запас уверенности решения, наиболее значимые положительные признаки, наиболее значимые отрицательные признаки и перечень признаков, обеспечивших преимущество над альтернативным классом. Такой формат объяснения удобен для вывода в web-интерфейсе и не требует от оператора анализа всей внутренней структуры модели.

Апробация метода проводилась на типовых сценариях, охватывающих шесть классов инцидентов: подбор учетных данных, компрометацию учетной записи, повышение привилегий, несанкционированный доступ к ресурсу, эксфильтрацию данных и действия

внутреннего нарушителя. Для пограничных случаев метод позволяет показать не только итоговый класс, но и его ближайшую альтернативу, например различать сценарии компрометации учетной записи и несанкционированного доступа к ресурсу. Для сценариев с недостаточной уверенностью автоматически формируется объяснение причины передачи инцидента аналитику. Это повышает прозрачность работы системы и снижает риск некритичного доверия к автоматическому решению. Таким образом, предложенный метод ориентирован на решение прикладной задачи интерпретации результатов классификации инцидентов информационной безопасности. Данный подход обеспечивает лучшую пригодность для интеграции в СППР, поддерживает аудит решений и может служить основой для дальнейшего развития механизмов объяснения в системах реагирования на инциденты.

Список литературы:

1. Басыня Е. А. Интеллектуальная поддержка при принятии технических решений в информационной инфраструктуре предприятия //Защита информации. Инсайд. – 2020. – №. 4. – С. 68-73.
2. Guidotti R. et al. A survey of methods for explaining black box models //ACM computing surveys (CSUR). – 2018. – Т. 51. – №. 5. – С. 1-42.
3. Ribeiro M. T., Singh S., Guestrin C. " Why should i trust you?" Explaining the predictions of any classifier //Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. – 2016. – С. 1135-1144.
4. MITRE ATT&CK [Электронный ресурс]. URL: https://attack.mitre.org/ (дата обращения: 19.05.2026).
5. NIST SP 800-61 Rev. 2: Computer Security Incident Handling Guide [Электронный ресурс]. URL: https://doi.org/10.6028/NIST.SP.800-61r2 (дата обращения: 19.05.2026).

Афанасьев И.О., Овасапян Т.Д., Грибков Н.А.

Санкт-Петербургский политехнический университет Петра Великого, Санкт-Петербург

АВТОМАТИЗАЦИЯ ПРОВЕРКИ ЗАЩИЩЕННОСТИ SIM-КАРТ ОТ АТАКИ ПО ПОБОЧНОМУ КАНАЛУ

В настоящее время SIM-карты продолжают широко применяться в телекоммуникационных системах и обеспечивают выполнение критически важных операций, связанных с аутентификацией абонента и защитой пользовательских данных. Несмотря на развитие программных механизмов защиты, актуальной остается задача оценки устойчивости указанных устройств к атакам, использующим особенности их аппаратной реализации. К числу таких относятся атаки по побочному каналу, основанные на анализе физических параметров функционирования устройства, в частности профилей энергопотребления [1, 2]. Проверка защищенности SIM-карт от подобных атак представляет собой многоэтапный процесс, включающий получение профилей энергопотребления, их предварительную обработку, синхронизацию и последующий анализ [3]. При этом возможность практического применения таких методов в составе программно-аппаратных комплексов во многом определяется степенью автоматизации указанных этапов. Несмотря на то, что значительная часть процедур может быть реализована в автоматическом режиме, полная автоматизация проверки защищенности затрудняется необходимостью принятия экспертных решений при обработке измеренных

данных. Одним из ключевых этапов является выбор эталонного профиля энергопотребления, относительно которого выполняется синхронизация всей совокупности полученных профилей.

Корректность выбора эталонного профиля оказывает существенное влияние на результат последующего анализа [4]. Профиль, содержащий временное смещение, шумовые выбросы или иные артефакты измерения, может привести к ошибочной синхронизации данных [5]. В результате снижается достоверность оценки зависимости между измеренным и смоделированным энергопотреблением, что может сделать проведение атаки невозможным даже при достаточном объеме экспериментальных данных.

В рамках доклада рассматривается подход к автоматизации проверки защищенности SIM-карт от атак по побочному каналу за счет исключения ручного выбора эталонного профиля энергопотребления. Предлагаемый подход основан на применении методов машинного обучения для классификации измеренных профилей и определения возможности их использования в качестве эталонных [6]. Для этого формируется набор признаков, характеризующих различные свойства профиля, включая показатели наличия аномальных участков или возможного временного смещения профиля. На основе размеченного набора данных производится обучение модели машинного обучения, результатом работы которой является оценка пригодности профиля для использования в качестве эталона.

Интеграция указанного подхода в общий процесс анализа позволяет сформировать автоматизированную последовательность проверки защищенности: получение профилей энергопотребления, расчет признаков, выбор эталонного профиля, синхронизация данных, проведение анализа и формирование результата проверки безопасности. Это снижает влияние человеческого фактора, уменьшает трудоемкость подготовки данных и повышает воспроизводимость экспериментов. В докладе приводится описание предлагаемого подхода и оценивается применимость методов машинного обучения для построения полностью автоматизированного программно-аппаратного комплекса проверки защищенности SIM-карт.

Библиографический список:

1. Жуковский Е. В., Москвин Д. А., Зегжда Д. П. Обнаружение программно-аппаратных закладок на основе анализа энергопотребления оборудования компьютера //Методы и технические средства обеспечения безопасности информации. – 2015. – №. 24. – С. 76-77.
2. Mangard S., Oswald E., Popp T. Power analysis attacks: Revealing the secrets of smart cards. – Boston, MA : Springer US, 2007.
3. Devine C., Pedro M. S., Thillard A. A practical guide to differential power analysis of USIM cards //Symposium Sur La Sécurité Des Technologies de l'Information et Des Communications (SSTIC). – 2018.
4. Kulow A. et al. Finding the needle in the haystack: Metrics for best trace selection in unsupervised side-channel attacks on blinded RSA //IEEE Transactions on Information Forensics and Security. – 2021. – Т. 16. – С. 3254-3268.
5. Debande N. et al. "Re-synchronization by moments": An efficient solution to align Side-Channel traces //2011 IEEE International Workshop on Information Forensics and Security. – IEEE, 2011. – С. 1-6.
6. Hettwer B., Gehrer S., Güneysu T. Applications of machine learning techniques in side-channel attacks: a survey //Journal of Cryptographic Engineering. – 2020. – Т. 10. – №. 2. – С. 135-162.

Басыня Е.А., Карапетьянц М., Жуков И.Ю.
Национальный исследовательский ядерный университет «МИФИ», Москва

АЛГОРИТМ ОЦЕНКИ ТОЖДЕСТВЕННОСТИ И ПРОФИЛИРОВАНИЯ ИНФОРМАЦИОННЫХ РЕСУРСОВ В ОВЕРЛЕЙНЫХ СЕТЯХ НА БАЗЕ ЛУКОВОЙ МАРШРУТИЗАЦИИ

Существующие исследования сосредоточены на двух направлениях. Системы непрерывного мониторинга и тематической категоризации ресурсов [1] обеспечивают высокую охватность, но не формируют модели тождественности при смене сетевых адресов. Алгоритмы дедупликации и выявления имитирующих ресурсов [2, 3] показывают, что лишь около 6 % страниц публичных корпусов уникальны, однако анализ ограничен одной плоскостью представления — текстом, разметкой или сертификатами, что не позволяет устойчиво отождествлять ресурс при его миграции.

Пусть R — множество наблюдаемых информационных ресурсов оверлейной сети на базе луковой маршрутизации, представимое в виде $R = R^{\text{uniq}} \cup R^{\text{mir}} \cup R^{\text{mig}} \cup R^{\text{phish}} \cup R^{\text{noise}}$, где R^{uniq} — оригинальные ресурсы; R^{mir} — копии оригинала по иному адресу; R_{mig} — ресурсы по новому адресу после прекращения функционирования по-прежнему; R^{phish} — внешне схожие с оригиналом ресурсы, не принадлежащие тому же оператору; R^{noise} — нерелевантный шум.

Для каждого ресурса формируется блочная признаковая модель следующего вида:

$$M(R) = \langle H_{\text{sim}}(R), H_{\text{min}}(R), A(R), S(R) \rangle,$$

где $H_{\text{sim}}(R) \in \{0,1\}^{\text{b}}$ — компактная текстовая сигнатура длины $b = 64$, построенная по алгоритму нечеткого хеширования SimHash и используемая для кандидатного отбора по расстоянию Хэмминга d_{H}. $H_{\text{min}}(R) = \langle m_1(R),\ldots,m_{\text{k}}(R) \rangle$ — MinHash-сигнатура длины k, дающая несмещенную оценку коэффициента Жаккара. $A(R) = \{a_1,\ldots,a_{\text{m}}\}$ представляет собой множество устойчивых семантических артефактов ресурса, сходство по которому вычисляется в форме взвешенного коэффициента Жаккара:

$$S_{\text{art}}(R1, R2) = \frac{\sum_{a \in A(R1) \cap A(R2)} w(a)}{\sum_{a \in A(R1) \cup A(R2)} w(a)}, \text{ с весовой функцией } w(a) = \log\left(\frac{N}{df(a)}\right),$$

где N − мощность корпуса наблюдаемых ресурсов, $df(a)$ − число ресурсов, содержащих артефакт a, что обеспечивает автоматическое повышение информативности редко встречающихся артефактов в соответствии с IDF-логикой.

Последним компонентом является структурное представление ресурса $S(R)$, которое описывает ресурс в в виде дерева $\Gamma(R)$ объектной модели документа со сходством $S_{\text{struct}}(R1, R2) = 1 - \frac{TED(\Gamma(R1),\Gamma(R2))}{Z(R1,R2)}$. В текущей интерпретации TED − расстояние редактирования деревьев, нормировочный множитель определяется выражением, определяющим верхнюю оценку стоимости преобразования $Z(R_1, R_2) = c_{\text{del}}|V_1| + c_{\text{ins}}|V_2|$, где $c_{\text{del}}, c_{\text{ins}} > 0$, $|V_1|$ и $|V_2|$ − мощности множеств вершин деревьев $\Gamma(\text{R}_1)$ и $\Gamma(\text{R}_2)$, соответственно, в свою очередь c_{del} и c_{ins} определяют стоимости операций удаления и вставки узла.

Итоговая интегральная мера сходства задается выпуклой комбинацией компонент и имеет следующий вид:

$$S_{\text{total}}(R1, R2) = \alpha J_{\text{text}} + \beta S_{\text{art}} + \gamma S_{\text{struct}}, \alpha, \beta, \gamma \geq 0, \alpha + \beta + \gamma = 1, 0 \leq S_{\text{total}} \leq 1.$$

Кандидатный отбор задается условием $c \in C(R_{\text{old}}) \Leftrightarrow d_{\text{H}}(H_{\text{sim}}(R_{\text{old}}), H_{\text{sim}}(c)) \leq d_{\text{max}}$. Для отобранных кандидатов принадлежность к одному из классов определяется решающим правилом следующего вида:

$$Y(Rold,c) = \begin{cases} R_{mir}, d_h(R_{old}, c) \le d_{strict}, J_{text}(R_{old}, c) \ge J_{clone}, S_{total}(R_{old}, c) \ge S_{clone} \\ R_{mig}, J_{mig} \le J_{text}(R_{old}, c) \le J_{clone}, S_{mig} \le S_{total}(R_{old}, c) \le S_{clone} \\ R_{phish},\ S_{rel} \le S_{total}(R_{old}, c) \le S_{mig} \\ R_{noise}, \text{иначе} \end{cases}$$

Пороги по расстоянию Хэмминга обоснованы вероятностной моделью SimHash, в которой d_H между b-битными сигнатурами распределено биномиально с параметром $p(\rho) = \arccos(\rho)/\pi$, где ρ — косинусная мера сходства текстовых представлений. В случае $b = 64$ значение $d_{max} = 16$ соответствует области грубого текстового сходства ($\rho \approx 0{,}71$), достаточной для кандидатного отбора, а значение $d_{strict} = 4$ соответствует области почти полного совпадения ($\rho \approx 0{,}98$) и используется как критерий клонирования. Пороги $J_{clone} = 0{,}90$, $J_{mig} = 0{,}60$, $S_{clone} = 0{,}85$, $S_{mig} = 0{,}70$, $S_{rel} = 0{,}50$ приняты в соответствии с эмпирическими практиками выделения почти идентичных групп ресурсов оверлейных сетей при сходстве выше 0,9 [2, 3].

Экспериментальная апробация проведена на синтетических данных в составе виртуального испытательного стенда, воспроизводящего условия функционирования оверлейной сети на базе луковой маршрутизации. Сравнение проводилось с пятью альтернативными конфигурациями, представленными в табл. 1.

Таблица 1 — Сравнительная оценка эффективности альтернативных конфигураций

Конфигурация	**TP**	**FP**	**FN**	**Precision**	**Recall**	**F_1**
Только SimHash	1496	134	302	0,918	0,832	0,873
Только MinHash	1783	109	15	0,942	0,991	0,965
Только артефакты	1267	10	531	0,992	0,705	0,824
Только структура	1546	426	252	0,784	0,860	0,820
Текст + структурные признаки	1770	204	27	0,897	0,985	0,938
Предлагаемый алгоритм	1791	41	94	0,978	0,951	0,964

Предлагаемый алгоритм превосходит четыре из пяти альтернативных конфигураций по интегральному показателю F_1 и обеспечивает наиболее сбалансированное соотношение достоверности и полноты на эталонной выборке. Значение Precision предлагаемого алгоритма превышает аналогичный показатель ближайшей по F_1 конфигурации на MinHash, а число ложных срабатываний снижено более чем в 2,5 раза. В задачах оперативного обнаружения утечек защищаемой информации такое соотношение методологически предпочтительнее, поскольку цена ложного отождествления ресурса как клона существенно превышает цену пропуска отдельного кандидата на этапе мониторинга. Минимальное отставание по F_1 предположительно связано с фиксированным экспертным выбором весов α, β, γ. В дальнейшем планируется корректировка весовых коэффициентов и повышение интегральных показателей качества за счет применения методов машинного обучения.

Список литературы:

1. Pastor-Galindo J., Mármol F. G., Pérez G. M. On the gathering of Tor onion addresses //Future Generation Computer Systems. – 2023. – Т. 145. – С. 12-26.
2. Matic S., Kotzias P., Caballero J. Caronte: Detecting location leaks for deanonymizing tor hidden services //Proceedings of the 22nd ACM SIGSAC conference on computer and communications security. – 2015. – С. 1455-1466.
3. Yoon C., Kim K., Kim Y., Shin S., Son S. Doppelgängers on the Dark Web: A Large-scale Assessment on Phishing Hidden Web Services // Proceedings of the World Wide Web Conference (WWW). — 2019. — P. 2225–2235.

Басыня Е.А., Сапегин В.Ю., Жуков И.Ю.
Национальный исследовательский ядерный университет «МИФИ», Москва

МЕТОДИКА АДАПТИВНОЙ БАЛАНСИРОВКИ ТРАФИКА МЕЖДУ НЕСКОЛЬКИМИ УСТАНОВЛЕННЫМИ ТУННЕЛЯМИ

Современные защищенные коммуникации все чаще используют не один маршрут, а несколько заранее установленных туннелей или цепочек ретрансляции: Tor-circuit, VPN-over-VPN, прокси-каскады, контейнеризованные шлюзы. При статическом выборе маршрута потоки конкурируют за одну очередь, что повышает задержку интерактивных сервисов и снижает устойчивость при перегрузке. Поэтому актуальна методика, учитывающая состояние цепочек, класс трафика и компромисс между скоростью, задержкой, стоимостью канала и приватностью.

Смежные исследования. За последние пять лет близкие результаты получены в трех направлениях. Во-первых, в Tor внедрены congestion control и Conflux/traffic splitting: несколько цепочек используются для повышения скорости и стабильности соединений, а решения о передаче опираются на задержку, признаки перегрузки и доступность цепочек. Во-вторых, в MPTCP и 5G ATSSS развиваются планировщики подпотоков, учитывающие RTT, доступную полосу, стоимость канала, буферы и класс приложения; это важно для web-, streaming- и video-on-demand-сценариев [1]. В-третьих, современные методы классификации зашифрованного трафика используют статистику потоков и машинное обучение, что позволяет различать типы приложений без раскрытия содержимого [2]. В отечественной литературе терминологические и прикладные основы виртуальных защищенных каналов связи, включая вопросы адаптивного управления защищенной коммуникационной инфраструктурой, последовательно рассматриваются в работах научной школы под руководством Е. А. Басыни [3].

Цель работы - разработать методику адаптивной балансировки сетевых потоков между несколькими установленными туннелями с учетом класса приложения и текущих параметров цепочек. Под балансировкой понимается выбор цепочки для нового потока, перераспределение долгоживущих передач и ограниченное дублирование критических запросов при деградации канала.

Предлагаемая методика включает четыре этапа. На первом этапе менеджер туннелей поддерживает пул активных цепочек и периодически измеряет для каждой из них вектор состояния: сглаженную RTT, джиттер, долю потерь/ретрансляций, текущую очередь, оценку доступной полосы, частоту сбоев, стоимость использования и уровень доверия к маршруту. Сглаженная RTT оценивается экспоненциальным скользящим средним, джиттер - как сглаженное среднее модуля ее приращения, занятость очереди - по соотношению фактической скорости доставки и темпа отправки, а доступная полоса - выборкой скорости доставки и методом парных пакетов. Каждая метрика имеет свою постоянную времени: быстрые (RTT, очередь) обновляются на масштабе секунд, медленные (стоимость, доверие) пересматриваются реже, что дает устойчивую к шуму, но реактивную картину пула. Измерения выполняются пассивно по пользовательским потокам и активными малыми пробами, причем суммарный бюджет проб ограничен фиксированной долей трафика и снижается при росте загрузки канала, чтобы не создавать заметной дополнительной нагрузки.

На втором этапе выполняется классификация потоков без анализа содержимого: используются порт и протокол, длительность соединения, распределение размеров пакетов, межпакетные интервалы, направление передачи, SNI/ALPN при наличии политики и локальная метка приложения. Предварительная классификация выполняется по первым пакетам соединения (правиловая эвристика по порту и форме первого обмена) и уточняется легкой моделью на потоковых признаках по мере накопления статистики; содержимое полезной нагрузки не анализируется, что сохраняет свойства приватности туннеля.

Выделяются классы «web-серфинг», «чат/мессенджер», «стриминг», «загрузка файлов», «фоновые обновления» и «служебный трафик». Для каждого класса задается профиль весов: web и чат минимизируют задержку, стриминг - стабильную полосу, загрузки - пропускную способность, фоновые потоки - стоимость.

На третьем этапе рассчитывается интегральная оценка $Score(c, k) = \Sigma\, w_i(k)\cdot m_i(c) - \Sigma\, \lambda_j\cdot p_j(c, k)$, где c - цепочка, k - класс трафика: взвешенная сумма нормированных метрик m_i за вычетом штрафов p_j за перегрузку очереди, недавние ошибки, частые переключения и превышение доли трафика на одной цепочке. Метрики приводятся к диапазону [0, 1] по текущему пулу цепочек: для показателей «чем меньше, тем лучше» (RTT, джиттер, потери, очередь, стоимость) нормировка инвертируется, для полосы и доверия применяется прямо. Весовые коэффициенты $w_i(k)$ задаются профилем класса экспертно или подстраиваются по скользящему окну; для исключения колебаний применяется гистерезис: поток переключается на другую цепочку лишь тогда, когда выигрыш в оценке устойчиво превышает порог на окне наблюдения.

На четвертом этапе выполняется назначение и перераспределение потоков. Короткие интерактивные соединения направляются в цепочку с максимальной оценкой. Длинные передачи могут дробиться на подпотоки, распределяемые по нескольким цепочкам пропорционально их оценкам, либо переключаться между цепочками на границах логических блоков, чтобы уменьшить out-of-order; привязка переключений к границам блоков вместе с буфером переупорядочивания на приемной стороне снижает долю пакетов, приходящих не по порядку, и связанные с этим повторные передачи. При деградации цепочки новые потоки туда не назначаются, а существующие получают ограничение скорости или мягкую миграцию без разрыва соединения; для критических служебных запросов в момент деградации допускается ограниченное дублирование по двум цепочкам с принятием первого пришедшего ответа, что повышает устойчивость при контролируемом росте нагрузки.

Критериями эффективности являются средняя и 95-процентильная задержка интерактивных запросов, время загрузки web-страниц, устойчивость битрейта стриминга, суммарная пропускная способность, потери, расход пробного трафика и равномерность нагрузки. Ожидаемый результат - снижение задержки web/чат-трафика при сохранении высокой утилизации каналов. Научная новизна состоит в объединении multipath-планирования, QoS-классификации и политики приватности на уровне клиента защищенных туннелей.

Таким образом, предложенная методика объединяет мониторинг цепочек, классификацию трафика и расчет интегральной оценки в единый замкнутый контур управления на стороне клиента, направляя каждый поток по наиболее подходящему туннелю без раскрытия содержимого и без доверенной координации с ретрансляторами. Дальнейшая работа предполагает экспериментальную оценку методики на реальных цепочках и подбор весовых профилей и порогов гистерезиса под конкретные классы приложений.

Список источников:

1. Amend M., Rakocevic V. Cost-efficient multipath scheduling of video-on-demand traffic for the 5G ATSSS splitting function // Computer Networks. – 2024. – Vol. 242. – Article 110218. – DOI: 10.1016/j.comnet.2024.110218.
2. Liu Z. et al. Fine-Grained Encrypted Traffic Classification Using Dual Embedding Mechanisms and Graph Neural Networks // Electronics. – 2025. – Vol. 14, № 4. – Article 778.
3. Басыня Е. А. Система самоорганизующегося виртуального защищенного канала связи // Защита информации. Инсайд. – 2018. – № 5. – С. 10–15.

Богина В.М., Дахнович А.Д.
Санкт-Петербургский политехнический университет Петра Великого, Санкт-Петербург

ПРИМЕНЕНИЕ ОНТОЛОГИЙ В ЗАДАЧЕ АНАЛИЗА НЕСТРУКТУРИРОВАННЫХ ТЕКСТОВЫХ ДАННЫХ С ПОМОЩЬЮ БЯМ

Стремительный рост объема неструктурированных данных в цифровой среде выступает одним из ключевых технологических вызовов современности. Согласно прогнозам IDC, к 2025 году общий объем мировых данных достиг 175 зеттабайт, из которых 80% приходится на неструктурированную информацию [1]. Вместе с ростом объемов неструктурированных данных рынок технологий обработки и анализа текстов демонстрирует устойчивую положительную динамику: объем глобального рынка технологий по обработке естественного языка (Natural Language Processing, NLP), выступающий технологической основой для извлечения полезной информации из текстов, достиг в 2025 году объема в 38,3 млрд долларов США при еще более высоком показателе CAGR — 30,9%[2].

Ввиду растущих объемов более актуальной становится задача извлечения полезной информации (знаний) из неструктурированного текста, сопряженная с рядом фундаментальных проблем, ограничивающих практическую применимость существующих решений. Большие языковые модели (БЯМ) склонны к генерации правдоподобного, но недостоверного ответа (галлюцинациям), что является критическим недостатком для систем прогнозной аналитики на текстовых данных, требующих строгого сохранения фактологии. Кроме того, БЯМ, являясь наиболее перспективным из существующих решений по анализу естественного языка, испытывают значительные трудности при формировании структурированного вывода с сохранением точности в задачах извлечения информации без дообучения, включая распознавание именованных сущностей (Named Entity Recognition, NER) и выделение отношений (relationship extraction, RE) [3], тогда как дообучение БЯМ наносит ограничения на обобщение и интерпретируемость получаемых результатов. Неточность, неполнота и недостоверность извлекаемых знаний из больших объемов неструктурированных текстовых данных формируют актуальную проблему, требующую разработки методов повышения качества автоматизированного анализа текстов.

Применение онтологий для извлечения знаний из неструктурированных текстовых данных способствует решению проблемы сохранения релевантной информации из больших объемов документов и повышению точности генерации ответов [4]. Однако, практическое применение онтологий сталкивается с серьезными ограничениями: их экспертное создание является длительным и дорогостоящим процессом; корректность автоматизированно созданных онтологий сложно доказуема [5]; универсальные относительно различных предметных областей онтологии представляют особую редкость, а объединение нескольких разрозненных требует большой аналитической работы по разрешению конфликтов между ними [6]. При углублении в конкретную предметную область возрастает потребность в детализации онтологического представления и большом количестве моделируемых показателей и связи между ними, что усиливает влияние описанных проблем.

В рамках настоящего исследования оценивается применимость широкой кросс-предметной онтологии Wikidata [7] для обработки узкоспециализированных текстов экономической направленности, собранных из тематических новостных каналов

мессенджера Telegram. Спецификой данного набора данных является наличие узкопредметной терминологии, привязка текстов ко времени публикации, большое количество обновляемых в контексте численных показателей и стилистические особенности кратких новостных текстов без подробного контекста. Применимость кросс-предметной онтологии Wikidata в рамках данного исследования оценивается в сравнении с узкоспециализированной экономической онтологией FIBO [8] и векторной реализацией RAG (Retrieval Augmented Generation) как базового решения на предмет объемов сохраняемой полезной информации и фактологическую точность генерации на ее основе. Валидация результатов проводится на размеченных в ручном режиме данных с помощью количественных и качественных показателей.

Список источников:

1. Global Unstructured Data Processing Software Market Research Report 2025 / QYResearch. — Electronic text data. — [S. l.]: QYResearch, 2025. URL: https://www.qyresearch.com/reports/4928366/unstructured-data-processing-software (дата обращения: 20.05.2026).
2. Natural Language Processing (NLP) Market Report 2026 / The Business Research Company. — Electronic text data. — [S. l.]: Research and Markets, 2026. — 250 p. URL: https://www.researchandmarkets.com/reports/5767538/natural-language-processing-nlp-market-report (дата обращения: 20.05.2026).
3. Yuzhang Xie. ACL Anthology / Yuzhang Xie. — Electronic text data. — [S. l.] : Association for Computational Linguistics, 2025. URL: https://aclanthology.org/people/yuzhang-xie/unverified/ (дата обращения: 20.05.2026).
4. Babaei Giglou H. et al. Beyond Statistical Parroting: Hard-Coding Truth into LLMs via Ontologies / Babaei Giglou H., Burbach S., Compagno F., Ismail M., Singh G., Rudolph S. – Proceedings of the Second International Workshop on Retrieval-Augmented Generation Enabled by Knowledge Graphs (RAGE-KG 2025), co-located with the 24th International Semantic Web Conference (ISWC 2025). — 2025. — 8 p.
5. A. S. Lippolis. Large Language Models Assisting Ontology Evaluation / A. S. Lippolis, M. J. Saeedizade, R. Keskisärkkä [et al.]. — Electronic text data. — [S. l.]: arXiv, 2025. URL: https://iris.cnr.it/handle/20.500.14243/559844 (дата обращения: 20.05.2026).
6. Osman I. et al. Ontology Integration: Approaches and Challenging Issues / Ben Yahia S., Diallo G. – Information Fusion. — 2021. — Vol. 71. — P. 38-63. — DOI: 10.1016/j.inffus.2021.01.007.
7. Wikidata: Main Page [Электронный ресурс] // Wikimedia Foundation. — Electronic text data. — [S. l.]:Wikidata. URL: https://www.wikidata.org/wiki/Wikidata:Main_Page (дата обращения: 20.05.2026).
8. FIBO Ontology [Электронный ресурс] / EDM Council. — Electronic text data. — [S. l.]: EDM Council. URL: https://spec.edmcouncil.org/fibo/ontology (дата обращения: 20.05.2026).

Болокан А.В., Зегжда Д.П.
Санкт -Пет ербургский полит ехнический университ ет Пет ра Великого, Санкт-Петербург

МУЛЬТИКАНАЛЬНЫЕ ИНФОРМАЦИОННЫЕ ВЕКТОРЫ ВОЗДЕЙСТВИЯ НА МОБИЛЬНЫЕ УСТРОЙСТВА: ТАКСОНОМИЯ И МЕТОДЫ ОБНАРУЖЕНИЯ

Современное мобильное устройство выступает не только средством связи, но и постоянно включенным вычислительным узлом, связанным с сотовой сетью, локальными радио-интерфейсами, приложениями, облачными сервисами, сенсорами и пользовательскими интерфейсами. Поэтому воздействие на него не сводится к эксплуатации одной программной уязвимости: оно может начинаться в одном канале, развиваться через другой и завершаться изменением состояния устройства, получением данных или навязыванием действия пользователю.

Под мультиканальным информационным вектором воздействия предлагается понимать совокупность технических и организационно-поведенческих действий, при которых сообщение, команда, сигнал, вложение, разрешение приложения или интерфейсный стимул передаются к мобильному устройству через один или несколько каналов и вызывают изменение состояния программной, сетевой, сенсорной или пользовательской среды. Такое определение расширяет классическую трактовку канала связи и учитывает прикладной, сетевой, физический и семантический уровни воздействия.

Таксономию таких векторов целесообразно строить по нескольким независимым векторам. По физической природе канала выделяются радиоканалы дальнего и ближнего действия (GSM/LTE/5G, Wi-Fi, Bluetooth, NFC), проводные каналы (USB, периферийные интерфейсы), оптические и визуальные каналы (камера, экран), акустические каналы (микрофон, динамик), а также электромагнитные и иные побочные каналы. По логическому уровню воздействия выделяются каналы платформы и ОС, прикладные каналы (мессенджеры, push-уведомления, браузер, SDK), сетевые протоколы и пользовательский интерфейс. По роли в атаке канал может использоваться для доставки полезной нагрузки, активации функции, повышения доверия, обхода политики доступа, фильтрации данных или маскировки активности.

Существующие исследования мобильной безопасности показывают, что статический анализ приложений, анализ разрешений и динамическое наблюдение за поведением остаются базовыми источниками признаков [1-3]. В то же время работы по скрытым голосовым и оптическим воздействиям демонстрируют, что атака может использовать штатные сенсоры и физические свойства микрофонов без классического вредоносного кода [4, 5]. Это подтверждает необходимость перехода от изолированного анализа приложений к корреляции событий на нескольких уровнях устройства.

Для обнаружения мультиканальных векторов предлагается гибридный подход. Первый контур основан на сигнатурах и правилах: подозрительные сочетания разрешений, обращений к API, сетевых доменов, типов вложений, радио-сессий и активности сенсоров. Второй контур использует поведенческие метрики: частоту активации интерфейсов, временную связанность событий, отклонения сетевых потоков, нетипичные последовательности системных вызовов, аномалии аудио- и радиоспектра. Третий контур выполняет межканальную корреляцию, представляя действия устройства в виде графа событий, где узлами являются процессы, разрешения, сетевые сессии, сенсорные обращения и пользовательские действия, а ребрами – причинно-временные связи между ними.

Практическим результатом такого подхода является матрица «канал – сценарий воздействия – признаки – метод обнаружения – возможное противодействие». Она позволяет сопоставлять SMS, push-уведомления, веб-ссылки, мессенджеры, Bluetooth/NFC-события, акустические команды и сенсорные аномалии в единой модели. Для исследования это дает основу не только для описания известных угроз, но и для

прогнозирования новых. Так при появлении нового канала доставки или нового способа связывания каналов, они могут быть классифицированы по тем же признакам.

Таким образом, научная значимость рассматриваемого направления заключается в переходе от перечня частных атак к единой таксономической и признаковой модели. Методы обнаружения должны учитывать гетерогенность мобильных платформ и сочетать статический анализ, динамическую телеметрию, сетевой мониторинг, анализ сенсорных данных и графовую корреляцию событий. Такой подход решает задачу выявления способов воздействия на мобильные устройства по информационным каналам различного типа и создает основу для последующей экспериментальной апробации.

Библиографический список:

1. Enck W., Octeau D., McDaniel P., Chaudhuri S. A Study of Android Application Security // Proceedings of the 20th USENIX Security Symposium. San Francisco, 2011.
2. Felt A.P., Chin E., Hanna S., Song D., Wagner D. Android Permissions Demystified // Proceedings of the 18th ACM Conference on Computer and Communications Security (CCS 2011). 2011. DOI: 10.1145/2046707.2046779.
3. Zhou Y., Jiang X. Dissecting Android Malware: Characterization and Evolution // 2012 IEEE Symposium on Security and Privacy. 2012. DOI: 10.1109/SP.2012.16.
4. Zhang G., Yan C., Ji X., Zhang T., Zhang T., Xu W. DolphinAttack: Inaudible Voice Commands // Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (CCS 2017). 2017. DOI: 10.1145/3133956.3134052.

Бровко Е.В., Пилькевич С.В.

Военно-космическая академия имени А.Ф. Можайского, Санкт-Петербург

ВЫЯВЛЕНИЕ АТАКУЮЩЕЙ КОМАНДНО-УПРАВЛЯЮЩЕЙ ИНФРАСТРУКТУРЫ НА ОСНОВЕ АНАЛИЗА МЕТАДАННЫХ СЕТЕВОГО ТРАФИКА

Командно-управляющая инфраструктура (C2-инфраструктура) является одним из ключевых элементов современных целевых атак. Под C2 понимается совокупность сетевых узлов, каналов связи, протокольных механизмов и серверных компонентов, с помощью которых злоумышленник поддерживает взаимодействие со скомпрометированными системами. Через такие каналы вредоносное программное обеспечение может получать команды, передавать результаты выполнения действий, обновлять конфигурацию, загружать дополнительные модули и сохранять устойчивое присутствие внутри атакуемой инфраструктуры [1].

Актуальность выявления C2-инфраструктуры возрастает в связи с усложнением сетевой среды. Широкое использование облачных платформ, CDN, арендованных виртуальных серверов, динамических доменных имен и защищенных протоколов связи приводит к тому, что вредоносная активность все чаще маскируется под легитимный сетевой обмен. Для специалистов центров мониторинга информационной безопасности это создает практическую проблему: одиночные индикаторы компрометации, такие как IP-адрес, домен или хеш-файл, быстро устаревают и не всегда позволяют своевременно обнаружить продолжающуюся атаку.

В таких условиях особое значение приобретает анализ метаданных сетевого трафика. В отличие от анализа содержимого передаваемых данных, работа с метаданными не требует расшифрования полезной нагрузки и может применяться даже в тех сегментах сети, где просмотр содержимого трафика невозможен по техническим, организационным или правовым причинам. При этом служебные параметры соединений часто сохраняют

признаки используемого сетевого стека, особенностей серверной конфигурации и поведения C2-канала. К таким признакам относятся DNS-параметры, характеристики TLS-соединений, сведения о сертификатах X.509, сетевые отпечатки JA4S и JA4X509, временные интервалы между соединениями, направление обмена, продолжительность сессий и соотношение объемов переданных данных [2-4]. Примеры используемых групп метаданных приведены в таблице 1.

Таблица 1. Используемые группы метаданных сетевого трафика

Источник метаданных	**Примеры признаков**	**Частные решаемые задачи**
DNS	Домен, TTL, частота запросов	Поиск динамики и скрытых каналов
TLS	Версия протокола, наборы шифров, JA4S	Идентификация сетевого стека C2
X.509	Поля сертификата, срок действия, JA4X509	Выявление повторяемых признаков серверов
Сетевые сессии	Длительность, интервалы, объемы данных	Оценка устойчивых шаблонов связи

Целью данного исследования является разработка подхода к выявлению C2-инфраструктуры на основе анализа метаданных сетевого трафика, а также описание логики программного комплекса, предназначенного для обнаружения подозрительных узлов, доменных имен и соединений по совокупности сетевых признаков. В рамках работы под программным комплексом понимается аналитическая система, которая принимает на вход PCAP-файл или поток сетевой телеметрии, извлекает из него значимые признаки, формирует профиль сетевого взаимодействия и оценивает степень его сходства с типовыми проявлениями командно-управляющих каналов.

Предлагаемый программный комплекс ориентирован на использование в составе средств мониторинга и реагирования на инциденты. Он не заменяет SIEM, IDS или EDR-системы, а дополняет их отдельным аналитическим контуром, позволяющим выявлять устойчивые сетевые признаки подозрительного взаимодействия и на их основе осуществлять: приоритезацию событий для аналитика; выделение наиболее подозрительных узлов; обнаружение дополнительных улик при расследовании инцидентов.

Схема работы комплекса основана на пассивном сборе и последующей обработке сетевой телеметрии. На первом этапе выполняется прием исходных данных: это может быть ранее сохраненный PCAP-файл, поток сетевого или результаты работы сенсоров сетевого мониторинга. На втором этапе осуществляется разбор протоколов и выделение параметров DNS, TLS, HTTPS и X.509. На третьем этапе полученные данные нормализуются и объединяются в профиль соединения [3]. Такой профиль включает криптографические, протокольные и поведенческие признаки, которые затем используются для корреляционного анализа.

Сетевой профиль соединения должен учитывать не только отдельные значения, но и контекст их появления. Например, сам по себе короткий срок действия сертификата, редкий домен или небольшой объем передаваемых данных не являются достаточным основанием для вывода о наличии C2-канала. Однако сочетание повторяющихся DNS-запросов, нестандартных параметров TLS-согласования, схожих отпечатков серверной стороны, регулярных обращений с фиксированными интервалами и асимметричного обмена данными может указывать на наличие скрытого командно-управляющего взаимодействия.

Результатом работы комплекса является перечень сетевых узлов, доменных имен, сертификатов и соединений, ранжированных по степени подозрительности. Каждому объекту может присваиваться интегральная оценка, отражающая количество и значимость выявленных признаков.

Архитектура программного комплекса может включать несколько функциональных модулей. Модуль приема трафика отвечает за загрузку PCAP-файлов или получение сетевой телеметрии в режиме, близком к реальному времени. Модуль разбора протоколов извлекает DNS-, TLS-, HTTPS- и X.509-параметры. Модуль нормализации приводит признаки к единому формату и устраняет дублирование. Хранилище признаков обеспечивает накопление профилей соединений и возможность ретроспективного анализа. Модуль корреляционного анализа выполняет сопоставление профилей с правилами и эталонными описаниями. Пользовательский веб-интерфейс отображает результаты анализа, уровень подозрительности, связанные события и объяснение причин срабатывания. При необходимости комплекс может дополняться внешними источниками данных.

Таким образом, выявление C2-инфраструктуры на основе анализа метаданных сетевого трафика является перспективным направлением повышения эффективности мероприятий информационной безопасности. Предложенный подход объединяет анализ DNS, TLS, X.509 и поведенческих характеристик сетевых сессий, что позволяет формировать устойчивый профиль подозрительного взаимодействия.

Список литературы:

1. MITRE ATT&CK. Command and Control, Tactic TA0011. URL: https://attack.mitre.org/tactics/TA0011/ (дата обращения: 26.04.2026).
2. FoxIO. JA4+ Network Fingerprinting. URL: https://github.com/FoxIO-LLC/ja4 (дата обращения: 26.04.2026).
3. Rescorla E. RFC 8446. The Transport Layer Security (TLS) Protocol Version 1.3. URL: https://www.rfc-editor.org/rfc/rfc8446 (дата обращения: 26.04.2026).
4. Cooper D., Santesson S., Farrell S., Boeyen S., Housley R., Polk W. RFC 5280. Internet X.509 Public Key Infrastructure Certificate and Certificate Revocation List Profile. URL: https://www.rfc-editor.org/rfc/rfc5280 (дата обращения: 26.04.2026).

Веденский В.А.

Санкт-Петербургский политехнический университет Петра Великого, Санкт-Петербург

МИНИМИЗАЦИЯ КОСВЕННЫХ СВЯЗЕЙ В КОНТЕЙНЕРНЫХ ОРКЕСТРАТОРАХ НА ОСНОВЕ РИСК-ОРИЕНТИРОВАННОГО ПЛАНИРОВАНИЯ

Контейнеризация и оркестрация на базе Kubernetes стали основой современных облачных сред, однако разделение ядра операционной системы между контейнерами порождает риск атак на совместное размещение [1,2]. Злоумышленник, скомпрометировавший один под, может через уязвимости ядра или побочные каналы воздействовать на соседние поды на том же узле [1]. Кроме того, горизонтальное перемещение атаки может осуществляться через сетевые соединения [3] или посредством кражи токенов ServiceAccount, смонтированных в файловую систему подов [4].

Существующие стратегии планирования либо вообще не учитывают безопасность (Spread, BinPack, Random), либо фокусируются на одном частном канале: например, минимизируют количество системных вызовов на узле (SySched [5], SmCPA [6]) или расхождение ролевой модели управления доступом (RBAClock [4]). Автором впервые предложена единая модель, охватывающая одновременно все три основных канала косвенных связей: «под – нода – под» через системные вызовы и через RBAC, а также «нода – под – нода» через сетевые взаимодействия. На основе модели формулируется задача оптимизации размещения подов и разрабатывается жадный алгоритм ее решения.

Формальное описание модели косвенных связей в контейнерном кластере:

1. Пусть заданы множества узлов $N = \{n_1, \dots, n_m\}$ с ресурсной емкостью Ω_j и подов $P = \{p_1, \dots, p_k\}$ с запросами ресурсов ω_i. Каждый под характеризуется профилем системных вызовов Σi (полученным динамическим анализом [5]) и бинарным вектором RBAC-привилегий $v_i = \{v_{i1}, \dots, v_{iR}\}$, Каждой привилегии сопоставлен вес критичности $w_r > 0$, определяемый по шкалам критичности облачных провайдеров или по внутренним политикам. Сетевое взаимодействие между подами описывается как $A = [a_{ik}]$ – матрицей сетевой связности; $a_{ik} \geq 0$ – интенсивность соединений. Переменная решения – матрица размещения $X = [x_{ij}]$, где $x_{ij} \in \{1{,}0\}$ $x_{ij} = 1$, если под p_i назначен на узел n_j, иначе – 0. Обязательны ограничения: $\sum_{\mathrm{i}} x_{ij} = 1,\ \forall i;$ – каждый контейнер в единственном экземпляре расположен на единственной ноде; $\sum_{\mathrm{i}} x_{ij}\omega_i \leq \Omega_j,\ \forall \mathrm{j}$ – ресурсные ограничения пода.

2. Косвенное взаимодействие «под – нода – под» может происходить по двум каналам: через использования уязвимостей общего ядра и расхождение RBAC-привилегий.

2.1. Риск ноды к эксплуатации уязвимостей общего ядра рассчитывается как взвешенная сумма всех используемых подами системных вызовов на этом узле. $\mathrm{R_{sys}(n_j)} = \sum_{\mathrm{s}\in \mathrm{U_{i:x_{ij}=1}}\sum \mathrm{i}} danger(\mathrm{s})$, где $danger(\mathrm{s})$>0 оценка опасности системного вызова (на основе CVE, категорий или экспертных весов). Чем меньше и безопаснее множество вызовов на узле, тем меньше вероятность успешной эксплуатации уязвимостей ядра [5].

2.2. Риск ноды n_j к расхождению RBAC-привилегий определим как сумму весов привилегий, присутствующих хотя бы у одного пода на узле: $R_{rbac}(n_j) = \sum_{r=1}^{R} w_r\, I[\bigvee_{k:x_{ki}=1} v_{kr} = 1]$. Такая метрика не поощряет совместное размещение подов с сильно различающимися уровнями прав.

3. Риск горизонтального перемещения при косвенном взаимодействие «нода – под – нода» происходит при сетевом взаимодействии двух подов на разных узлах и рассчитывается по формуле $\mathrm{R_{net}(n_j, n_l)} = \sum_{i:x_{ij}=1;k:x_{kl}} a_{ik}$.

4. Для расчета ресурсного дисбаланса используется дисперсия [6] $\eta = \frac{1}{|N|}\sum_j (u_j - u')^2$, где $u_j = \frac{1}{\Omega_j}\sum_i x_{ij}\,\omega_i$, $u' = \frac{1}{|N|}\sum_j u_j$.

5. Целевая функция представляет собой линейную комбинацию рисков и дисбаланса: $\min_X F = \alpha \sum_j \mathrm{R_{sys}(n_j)} + \beta \sum_j \mathrm{R_{rbac}(n_j)} + \delta \sum_j \mathrm{R_{net}(n_j, n_l)} + \gamma\eta$ при ограничениях $\sum_{\mathrm{i}} x_{ij} = 1,\ \forall i;\ \sum_{\mathrm{i}} x_{ij}\omega_i \leq \Omega_j,\ \forall \mathrm{j}$ Веса αβγδ задают приоритеты. Их выбор может осуществляться экспертами или методами многокритериальной оптимизации (например, исследованием Парето-фронта). Задача содержит нелинейные комбинаторные зависимости и на практике труднорешаема для кластеров среднего и большого размера, что оправдывает применение эвристик.

Формально определен алгоритм на базе жадной стратегии последовательного назначения подов с выбором узла, дающего минимальное приращение целевой функции:

1. Ранжирование подов по уровню критичности. $criticality(p_i) = \mathrm{c_1}\sum_k a_{ik} + \mathrm{c_2}\max_{s\in\Sigma i} danger(s) + \mathrm{c_3}(v_{i1} * W)$.

2. Размещение:

2.1. Инициализация: x_{ij}=0; для каждого узла хранятся текущие объединенные профили вызовов $\Sigma^{(j)}$, индикаторы привилегий $V_r^{(j)}$ и остаточная емкость.

2.2. Для очередного пода p_i и каждого допустимого узла n_j вычисляется приращение ΔF_{ij}.

2.3. Под назначается на узел с минимальным ΔF_{ij}.

2.4. Обновляется состояние узла.

Разработанный метод имеет ряд ограничений. Модель статична и требует априорного знания профилей системных вызовов и RBAC, не учитывает динамические изменения

кластера, в т.ч. масштабирование и миграцию подов. Веса *danger(s)* и привилегий w_r основаны на экспертных оценках, что может влиять на качество защиты. Жадный алгоритм не гарантирует глобальный оптимум, хотя его оперативность делает его применимым в практических планировщиках.

Таким образом, в ходе исследования построена формальная модель трехканального косвенного взаимодействия «под – нода – под» и «нода – под – нода» в Kubernetes-кластере, сформулирована задача минимизации нормированного составного риска и предложен алгоритм минимизации риска. Дальнейшая работа предполагает экспериментальную валидацию на реалистичных кластерных конфигурациях и расширение на случай динамического изменения состава подов.

Список литературы:

1. Shringarputale S. Et al. Co-residency Attacks on Containers are Real // CCSW 2020 - Proceedings of the 2020 ACM SIGSAC Conference on Cloud Computing Security Workshop. Association for Computing Machinery, Inc, 2020. P. 53-66.
2. Gao X. et al. ContainerLeaks: Emerging Security Threats of Information Leakages in Container Clouds // Proceedings - 47th Annual IEEE/IFIP International Conference on Dependable Systems and Networks, DSN 2017. Institute of Electrical and Electronics Engineers Inc., 2017. P. 237-248.
3. Santos J. et al. Towards Intent-Based Scheduling for Performance and Security in Edge-to-Cloud Networks // 2024 27th Conference on Innovation in Clouds, Internet and Networks (ICIN). IEEE, 2024. P. 222-227.
4. Chen Q. et al. RBAClock: Contain RBAC Permissions through Secure Scheduling // Proceedings - 28th International Symposium on Research in Attacks, Intrusions and Defenses, RAID 2025. Institute of Electrical and Electronics Engineers Inc., 2025. P. 950-965.
5. Le M. V. et al. Securing Container-based Clouds with Syscall-aware Scheduling // Proceedings of the ACM Conference on Computer and Communications Security. Association for Computing Machinery, 2023. P. 812-826.
6. Li M., Hu H., Liu W. A Secure Container Placement Algorithm Based on Microservice Invocation Criticality // 2024 IEEE 2nd International Conference on Control, Electronics and Computer Technology, ICCECT 2024. Institute of Electrical and Electronics Engineers Inc., 2024. P. 234-240.

Горецкий И.А., Полтавцева М. А.

Санкт-Петербургский политехнический университет Петра Великого, Санкт-Петербург

МЕТОД ВЫБОРА СОВМЕСТИМЫХ КОНТРМЕР ДЛЯ ПРОТИВОДЕЙСТВИЯ АТАКАМ В ПРОГРАММНО-ОПРЕДЕЛЯЕМЫХ СЕТЯХ (SDN)

Программно-определяемые сети (Software-Defined Networking, SDN) являются одним из перспективных подходов к построению управляемых и динамически реконфигурируемых сетевых инфраструктур. Ключевой особенностью SDN является разделение плоскости управления, передачи данных и приложений, что позволяет централизованно изменять правила маршрутизации и обработки сетевого трафика. Данная особенность повышает гибкость управления сетью, однако одновременно формирует новые угрозы безопасности, связанные с атаками на контроллер, каналы управления, сетевые приложения и таблицы потоков коммутаторов [1].

Одной из актуальных задач обеспечения безопасности SDN является выбор эффективного набора мер противодействия сетевым атакам. В существующих работах

рассматриваются различные подходы к обнаружению и нейтрализации атак, включая фильтрацию вредоносного трафика, перенаправление потоков, ограничение скорости передачи, изоляцию скомпрометированных узлов и изменение правил OpenFlow [2]. Однако применение таких мер может приводить к дополнительной нагрузке на контроллер, увеличению задержек, конфликтам между правилами потоков и нарушению доступности легитимных сервисов.

В данной работе предлагается метод выбора совместимых контрмер для противодействия атакам в программно-определяемых сетях. Основная идея метода заключается в том, чтобы не рассматривать всю сеть целиком, а сначала локализовать область воздействия атаки и сформировать подграф, включающий только затронутые узлы, каналы связи и потоки. За счёт этого сокращается пространство возможных состояний сети, а также уменьшается количество комбинаций контрмер, которые необходимо анализировать.

Формально SDN-сеть может быть представлена в виде графа $\mathrm{G} = (V, E)$, где V - множество сетевых узлов, а E – множество каналов связи между ними. После обнаружения атаки выделяется локальный подграф $G_{local}(V_{local}, E_{local})$, описывающий область сети, на которую оказывает влияние атака. Далее для данного подграфа формируется множество допустимых контрмер $C = \{c_1, c_2, \dots, c_n\}$, каждая из которых рассматривается как преобразование текущего состояния сети.

Предлагаемый метод состоит из следующих этапов:

1. Формирование локального подграфа, соответствующего области воздействия атаки.
2. Построение множества возможных контрмер, применимых к данной области.
3. Формальное описание контрмер как преобразований над графом сети.
4. Оценка совместимости контрмер с учетом конфликтов правил, задержек, загрузки каналов и ресурсов контроллера.
5. Выбор набора контрмер, обеспечивающего минимизацию ущерба от атаки при сохранении работоспособности сети.

Особенностью метода является анализ совместимости контрмер. В SDN конфликтующие правила потоков и политики безопасности между разными плоскостями сети приводят к некорректной обработке трафика и открытию новых путей для атаки на сеть. Поэтому разрешения таких конфликтов является важной частью управления безопасностью сети [3]. В предлагаемом подходе контрмеры рассматриваются не изолированно, а совместно, что позволяет заранее выявлять ситуации, когда одна мера снижает эффективность другой либо совместное применение мер создает чрезмерную нагрузку на инфраструктуру сети.

Предлагаемый метод направлен на повышение эффективности противодействия атакам в SDN за счет локализации области анализа и выбора совместимых мер защиты. Использование локального подграфа позволяет уменьшить вычислительную сложность задачи, а формализация контрмер создает основу для автоматизированного выбора защитных действий в условиях динамически изменяющейся сетевой среды.

Список литературы:

1. Alsmadi I., Xu D. Security of software defined networks: A survey //Computers & security. – 2015. – Т. 53. – С. 79-108.
2. Eliyan L. F., Di Pietro R. DoS and DDoS attacks in Software Defined Networks: A survey of existing solutions and research challenges //Future Generation Computer Systems. – 2021. – Т. 122. – С. 149-171.
3. Wang J. et al. A method of OpenFlow-based real-time conflict detection and resolution for SDN access control policies //Chinese Journal of Computers. – 2015. – Т. 38. – №. 4. – С. 872-883.

Денисенко В.В., Борисенко В.В., Андрушкевич Д.В.
Военно-космическая академия имени А.Ф. Можайского, Санкт-Петербург

ПОДХОД К ВЫБОРУ МЕТОДА ЗАЩИТЫ ВЫЧИСЛИТЕЛЬНОЙ СЕТИ С УЧЕТОМ ЕЕ ТОПОЛОГИИ

Распространение сетевых червей приводит к последовательной компрометации узлов вычислительных сетей. Согласно Федеральному закону № 187-ФЗ и Стратегии национальной безопасности РФ, ключевым требованием к инфраструктуре является обеспечение ееустойчивого функционирования при компьютерных атаках (КА)[1, 2]. В современных условиях выполнение этого требования напрямую зависит от живучести сети – способности сохранять работоспособность и связность критического ядра даже при частичном компрометировании узлов. Традиционные методы защиты (сигнатурный анализ, эвристические правила, ручная настройка) часто игнорируют топологические свойства графа, что не позволяет эффективно обеспечивать живучесть сети [3]. В работе предложен подход к выбору метода защиты сети с учетом ее топологии, основанный на автоматизированном подборе алгоритма сегментации сети посредством количественного анализа графа [4].

Цель исследования –повышение живучести инфраструктуры за счет учета ее топологических характеристик при выборе оптимального метода сегментации сети.

Функциональная модель процесса выбора защиты разработана в нотации IDEF0, что позволило формализовать входные параметры (топология графа $G = (V, E)$), управляющие воздействия (алгоритмы сегментации), механизмы реализации (программные библиотеки NetworkX для графового анализа, SQLite для хранения результатов) и выходные результаты (рекомендации по выбору алгоритма защиты). Программный комплекс моделирует распространение червя по принципу обхода графа в ширину (BFS). Исследование проведено на эталонных топологиях («Звезда», «Кольцо», «Полносвязная», «Гибридная») масштабом до 500 узлов. Сравнительному анализу подверглись два метода защиты:

- алгоритм Гирвана–Ньюмана (выявление и разрыв связей между сообществами);
- BFS-кластеризация (локализация угрозы вокруг узлов с высокой центральностью).

Ключевой метрикой эффективности выступал параметр структурной живучести, определяемый как размер наибольшего связного компонента незараженных узлов (LCC). Данный показатель количественно отражает способность сети сохранять работоспособность критического ядра в условиях атаки.

Эксперименты показали, что адаптивный выбор метода защиты позволяет не только замедлить распространение атаки на 40–70%, но и сохранить связность критического ядра сети. Установлено, что:

- для полносвязных структур (Mesh) максимальную живучесть обеспечивает алгоритм Гирвана–Ньюмана, эффективно изолирующий зараженные кластеры;
- для централизованных топологий («Звезда») наиболее эффективна BFS-кластеризация, позволяющая локализовать инцидент при минимальном числе удаляемых связей;
- гибридные топологии требуют комбинированного подхода, что подтверждает необходимость предварительного структурного анализа перед выбором контрмеры [5].

Таким образом, предложенный подход повышает структурную живучесть вычислительной сети при последовательной компрометации узлов за счет учета ее топологических свойств при выборе метода защиты. Это обеспечивает гарантированное сохранение связности критического ядра инфраструктуры, минимизирует количество разрываемых связей и исключает ручной подбор параметров, снижая влияние человеческого фактора на процесс локализации компьютерных атак [6].

Дальнейшее развитие работы связано с интеграцией модели со средой эмуляции EVE-NG для оценивания влияния методов сегментации на производительность сети и задержки передачи данных в условиях реальной нагрузки.

Список литературы:

1. Федеральный закон от 26.07.2017 № 187-ФЗ «О безопасности критической информационной инфраструктуры Российской Федерации» // Собрание законодательства РФ. – 2017. – № 31. – Ст. 4736.
2. Указ Президента РФ от 02.07.2021 № 400 «О Стратегии национальной безопасности Российской Федерации» // Собрание законодательства РФ. – 2021. – № 27 (часть II). – Ст. 5351.
3. Хорошко В., Хохлачова Ю., Вишневская Н. Декомпозиция технологии компьютерных сетей при их проектировании // Научный журнал по информационной безопасности. – 2023. – Т. 29, N 3. – С. 130-137.
4. Лаврова Д.С., Зегжда Д.П., Зайцева Е.А. Моделирование сетевой инфраструктуры сложных объектов для решения задачи противодействия кибератакам // Вопросы кибербезопасности. – 2019. – N 2 (30). – С. 13-29.
5. Елисеев Е.Э., Бурлаков М.Е. Применение адаптивных алгоритмов в графовых структурах при решении задач предотвращения компьютерных угроз: выпускная квалификационная работа. – Самара: Самарский университет, 2021.
6. Al-Eiadeh M.R., Abdallah M. GeniGraph: A genetic-based novel security defense resource allocation method for interdependent systems modeled by attack graphs // Computers & Security. – 2024. – Vol. 144. – Art. 103927.

Добренков В.А., Теткин С.В., Киселёв А.Н.

Военно-космическая академия имени А.Ф. Можайского, Санкт-Петербург

РАЗРАБОТКА И ИССЛЕДОВАНИЕ МОДУЛЯ АНАЛИЗА СОДЕРЖИМОГО ФАЙЛОВ ДЛЯ DLP-СИСТЕМЫ МОНИТОРИНГА СЕТЕВОЙ АКТИВНОСТИ

Проблема предотвращения утечек конфиденциальной информации через сетевые каналы становится всё более острой в условиях цифровой трансформации организаций. Традиционные методы ручного анализа сетевого трафика и проверки передаваемых файлов, такие как выборочный аудит и пост-инцидентный анализ, становятся неэффективными из-за растущих объёмов данных и необходимости реагирования в реальном времени. В последние годы наблюдается значительный сдвиг в сторону автоматизированных DLP-решений с применением контентной фильтрации и анализа на основе правил [1, 5].

Особый интерес представляет подход, основанный на использовании модуля анализа содержимого файлов, встраиваемого в конвейер перехвата и анализа сетевого трафика. Этот метод демонстрирует существенные преимущества по сравнению с традиционными подходами благодаря способности в реальном времени извлекать текст из широкого спектра форматов файлов и применять контекстные правила обнаружения конфиденциальных данных.

Современные методы анализа содержимого файлов имеют ряд недостатков, таких как: ограниченная поддержка форматов файлов (особенно вложенных архивов и документов со сканированными изображениями), высокий уровень ложных срабатываний при использовании простых регулярных выражений, отсутствие устойчивости к базовым методам обфускации (вставка нулевых символов, разбиение цифр пробелами), недостаточная производительность для работы в режиме реального времени и сложность

интеграции с существующими сетевыми перехватчиками [1, 5]. Во многом это зависит от качества используемых инструментов для извлечения текстовой информации из различных источников, ограниченности типов анализируемых артефактов и отсутствия унифицированного конвейера обработки.

Решением этих проблем служит разработанный модуль анализа содержимого файлов для DLP-системы мониторинга сетевой активности. Подход включает конвейерную обработку: приём файлов из очереди (HTTP, SMTP, FTP), определение MIME-типа через сигнатурный анализ (libmagic), извлечение текста с помощью многоформатного экстрактора (DOCX, XLSX, PPTX, PDF, RTF, HTML, ODT, ZIP/RAR с рекурсивной распаковкой до 3 уровней, а также изображения JPEG, PNG, GIF, TIFF через OCR [3, 6]), нормализацию текста (удаление лишних пробелов, восстановление кодировки), применение движка правил (регулярные выражения для паспортов РФ, ИНН, номеров карт; словари маркеров конфиденциальности) с выдачей вердикта «блокировать/разрешить/карантин». Дополнительно модуль поддерживает асинхронную очередь с приоритетами (реального времени vs фоновые задачи) и многопоточную обработку для достижения целевой производительности. Определение формата файла происходит по сигнатурам (магическим числам), что делает анализ нечувствительным к изменению расширения файла [2].

Предложенный подход к анализу содержимого файлов на основе конвейерной архитектуры и гибридного движка правил показал высокую эффективность и перспективы для практического применения. Результаты экспериментов на корпусе из 5000 файлов подтверждают возможность использования данного метода в реальных условиях мониторинга сетевой активности организации. Достигнуты целевые показатели: полнота обнаружения (Recall) — 92%, точность (Precision) — 96%, задержка обработки файла объёмом до 10 МБ — 87 мс (p95), пропускная способность — 55 МБ/с на одном ядре CPU.

В будущем планируется реализовать: расширение набора поддерживаемых форматов файлов (включая MSG, EML, AutoCAD, табличные форматы аналитики [4]), улучшение обработки обфусцированных данных (автоматическое восстановление цифровых последовательностей), разработка механизмов объяснения принятых решений для расследований, интеграция дополнительных сканеров (YARA-правила, ML-классификаторы), поддержка отечественных форматов электронной подписи и реестров. Также планируется развитие модуля в сторону адаптивного выбора стратегии извлечения текста в зависимости от типа файла и доступных вычислительных ресурсов с использованием кэширования результатов OCR для повторяющихся изображений [6]. В будущем проектируемый модуль также будет предусматривать интеграцию с оркестраторами контейнеров и системами SIEM для централизованного мониторинга и корреляции событий. Это позволит замкнуть цикл защиты от момента перехвата сетевого пакета до вынесения решения о блокировке передачи и существенно повысить уровень защищённости информационной системы организации.

Данный подход может стать основой для создания эффективных систем предотвращения утечек информации (DLP) с функцией контентного анализа сетевого трафика, что особенно важно в условиях растущих требований к защите персональных данных и коммерческой тайны.

Список использованных источников:

1. Вахонин С. Эффективность применения контентной фильтрации в DLP-системах [Электронный ресурс]. URL: https://www.itweek.ru/security/article/detail.php?ID=181038
2. Stakhanovets. DLP — По форматам файлов [Электронный ресурс]. URL: https://help.stakhanovets.ru/spravka/dlp-po-formatam-fajlov-146/
3. Stakhanovets. DLP — Общие настройки [Электронный ресурс]. URL: https://help.stakhanovets.ru/spravka/64/

4. Netskope. Supported File Categories and File Types [Электронныйресурс]. URL: https://docs.netskope.com/en/supported-file-categories-and-file-types/
5. Вахонин С. Эффективность применения контентной фильтрации в DLP-системах [Электронный ресурс]. URL: http://itsec.ru/articles2/dlp/effektivnost-primeneniya-kontentnoy-filtratsii-v-dlp-sistemah
6. Stakhanovets. DLP — OCR [Электронный ресурс]. URL: https://help.stakhanovets.ru/spravka/147/

Дорошенко Д.С., Гильмуллин Р.М., Резанов Д.И.
Военно-космическая академия имени А.Ф. Можайского, Санкт-Петербург

ПРОГРАММНЫЙ КОМПЛЕКС ПОИСКА ИСКАЖЕННЫХ СИГНАТУР ФАЙЛОВ С ИСПОЛЬЗОВАНИЕМ РАССТОЯНИЯ ЛЕВЕНШТЕЙНА

В условиях стремительного роста объемов цифровой информации задача поиска скрытых, намеренно искаженных или случайно поврежденных данных в бинарных файлах приобретает критическое значение для специалистов в области цифровой форензики и анализа вредоносного программного обеспечения. Файлы, передаваемые по информационно-телекоммуникационным сетям, могут содержать стеганографически внедренные вложения, встроенные архивы с вредоносным кодом, а также фрагменты удаленных или поврежденных объектов [1]. Способность эффективно обнаруживать и извлекать такие данные является необходимым условием для проведения полноценной компьютерно-технической экспертизы.

Существующие инструменты для извлечения данных из бинарных файлов, такие как Binwalk, Foremost или Scalpel стали стандартом в рассматриваемой предметной области. Однако они ориентированы на выполнение в интерфейсе командной строки и используют алгоритмы поиска точных соответствий с предопределенной эталонной сигнатурой без учета возможных искажений [2]. Такие методы неэффективны в случаях, когда сигнатура файла была преднамеренно изменена злоумышленником путем модификации нескольких байт с целью сокрытия вредоносного кода.

Для преодоления указанных ограничений авторами статьи разработан программный комплекс, ключевой особенностью которого является реализация механизма нечетного поиска искаженных сигнатур с использованием расстояния Левенштейна [3].

Основой для идентификации файлов в программе служит база сигнатур, представляющая собой совокупность байтовых последовательностей, характерных для различных форматов данных. В общем виде процесс поиска можно описать следующим образом: для заданной сигнатуры $S = \{s_1, s_2, \dots, s_n\}$ и входного бинарного потока B = $\{b_1, b_2, ..., b_m\}$ производится поиск позиции i, такой что подпоследовательность B[i…i+n-1] = S для детерминированного поиска либо удовлетворяет условию d(B[i…i+n-1], S) ≤ T для нечеткого поиска, где d — функция расстояния, T — заданный порог.

При разработке архитектуры программного комплекса особое внимание уделено обеспечению возможности работы с файлами больших размеров, объем которых может превышать объем доступной оперативной памяти. Для решения данной задачи реализован механизм чанкового чтения, при котором файл обрабатывается последовательно блоками фиксированного размера. На рисунке 1 представлена структурная схема программного комплекса, отражающая основные вычислительные модули и их функции.

Расстояние Левенштейна между двумя строками, в данном случае байтовые последовательности, определяется как минимальное количество операций вставки, удаления и замены символов, необходимых для преобразования одной строки в другую [3]. В программном комплексе данная метрика применяется для сравнения эталонной сигнатуры файла с фрагментами анализируемых данных, при этом фиксируются те

позиции, где выявленные различия не превышают заданного пользователем порогового значения. Это позволяет обнаруживать файловые структуры с изменёнными заголовками, что существенно расширяет возможности анализа в условиях применения злоумышленниками методов обфускации и модификации файлов.

Для оценки эффективности разработанного программного комплекса была проведена серия экспериментов на тестовой выборке бинарных данных. Вероятность ложноотрицательных результатов в подобных задачах обусловлена объективными факторами, не зависящими от корректности реализации алгоритма. К ним относятся: смещение сигнатуры относительно границы блока при блочном чтении данных, особенности метрики Левенштейна при обработке последовательностей с повторяющимися байтами, а также структурные спецификации форматов файлов, при которых сигнатура частично перекрывается служебными данными или метаданными.

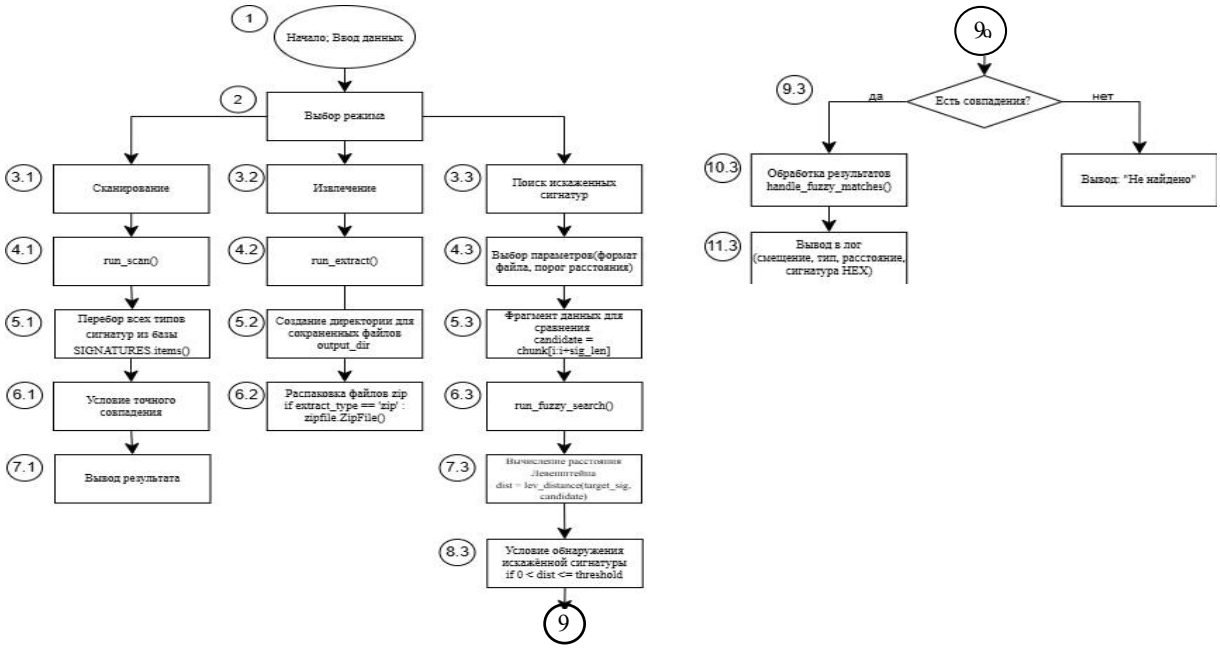


Рисунок 1 - Структурная схема программного комплекса

Экспериментальные результаты, отражающие зависимость доли пропусков от значения порога принятия решения и степени искажения данных, представлены в таблице 1.

Таблица 1 – Зависимость вероятности пропуска сигнатуры от порога принятия решения и числа модифицированных байтов

Значение порога	Максимальное число искажённых байтов при 0 % пропусков	Доля ложноотрицательных результатов при превышении порога искажений
2	≤ 2 байт	65 % (при 3 байтах)
4	≤ 4 байт	70 % (при 5–6 байтах)
5	≤ 5 байт	30 % (при 6 байтах)

Анализ полученных данных демонстрирует высокую эффективность предложенного подхода. Увеличение порога принятия решения позволяет сохранять нулевой уровень пропусков при большем количестве модификаций. Наилучший баланс между чувствительностью и устойчивостью к помехам обеспечивает порог, равный 5:

ложноотрицательные срабатывания фиксируются только при шести изменённых байтах и не превышают 30%, что гарантирует обнаружение не менее трети объектов даже при значительном повреждении их структуры.

Таким образом, даже с учетом объективных ограничений алгоритма, разработанный комплекс демонстрирует высокую вероятность обнаружения, а процент пропусков не влияет
на его практическое использование в задачах цифровой форензики.

Список литературы:

1. Федоров, Н.В. Компьютерная криминалистика: учебное пособие / Н.В. Федоров. – М.: НИУ ВШЭ, 2018. – 320 с.
2. Документация Binwalk [Электронный ресурс] // ReFirmLabs. – Режим доступа: https://github.com/ReFirmLabs/binwalk (дата обращения: 04.03.2026).
3. Левенштейн, В.И. Двоичные коды с исправлением выпадений, вставок и замещений символов / В.И. Левенштейн // Доклады Академии Наук СССР. – 1965. – Т. 163, № 4. – С. 845–848.

Ерастов В.О., Зубков Е.А., Зегжда Д.П.

КОМПЛЕКСНЫЙ ПОДХОД К ВЫЯВЛЕНИЮ ДОСТУПНЫХ IPV6-УЗЛОВ В СЕТИ ИНТЕРНЕТ

Исчерпание адресного пространства IPv4 и принципиальная невозможность его преодоления с помощью NAT и CIDR сделали переход на протокол IPv6 единственным долгосрочным решением. Однако внедрение IPv6 с его 128-битной адресацией и колоссальным адресным пространством породило новую исследовательскую проблему. Классические методы сканирования, основанные на полном переборе, здесь принципиально неработоспособны, а плотность активных узлов остаётся крайне низкой. Таким образом, обнаружение доступных IPv6-узлов в масштабах сети Интернет становится нетривиальной задачей, требующей принципиально иных подходов.

В рамках настоящей работы был разработан комплексный подход, объединяющий несколько взаимодополняющих методов:

1. Сужение диапазона сканирования. Анализ иерархической политики распределения адресных префиксов позволяет выделить реально используемые блоки: провайдерам выдаются префиксы /32, конечным пользователям – /48 или /56, а для конечной подсети стандартом закреплён префикс /64. Дальнейшее сокращение перебора достигается за счёт анализа типовых шаблонов формирования идентификатора интерфейса: ручное назначение адреса (::1, ::2:80, встраивание IPv4, осмысленные HEX-слова), автоконфигурация SLAAC на основе EUI-64 (где известны 24-битные OUI производителя и константа FF:FE в середине, что сокращает перебор до 2^{24} вариантов) и последовательная выдача адресов из пула с использованием протокола DHCPv6.
2. Ответное сканирование. В отличие от активного перебора, здесь адреса собираются пассивно, из входящего трафика легитимных сервисов. В частности, включение собственного сервера в проект NTP POOL позволяет регистрировать IPv6-адреса клиентов, обращающихся для синхронизации времени. Дополнительным источником адресов выступает протокол BitTorrent, спецификация которого предусматривает передачу списков пиров с их IP-адресами, включая IPv6.

3. Поиск по доменным именам. Домены, поддерживающие IPv6, выявляются как перебором по словарям, так и с помощью специализированных поисковых запросов с последующим разрешением AAAA-записей.

4. Применение модели машинного обучения в качестве генератора адресов-кандидатов. Модель обучается на корпусе адресов, накопленных пассивными методами (NTP, доменные имена), и формирует вероятностное распределение наиболее ожидаемых идентификаторов интерфейса и подсетей. Сгенерированные ML-моделью адреса поступают на активное сканирование, а подтверждённые как активные узлы возвращаются в обучающую выборку. Таким образом, реализуется замкнутый цикл: модель непрерывно дообучается по результатам каждого цикла сканирования, повышая точность предсказаний и последовательно расширяя покрытие обнаруживаемого адресного пространства.

Ключевой результат работы состоит в комбинировании этих методов. Даже при известном /32 префиксе полный перебор занял бы порядка 2,5 квадриллиона лет (при сканировании 10^6 узлов в секунду). Предложенный же подход – сбор реальных адресов, восстановление по ним шаблона генерации и последующее прицельное сканирование – кардинально сокращает пространство поиска. Например, при обнаружении шаблона встраивания IPv4 перебор в основном ограничивается 2^{16} адресами, а для SLAAC – 2^{40} адресами, что при использовании инструментов класса ZMap выполнимо приблизительно за 12 дней.

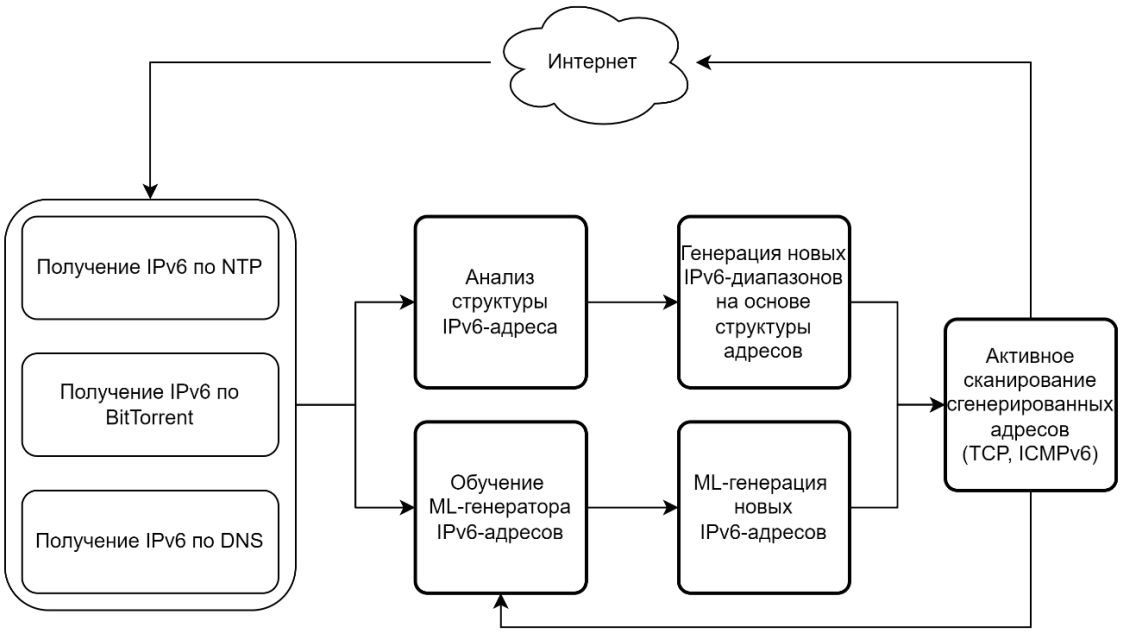


Рисунок 1 – Блок-схема подхода к выявлению доступных IPv6-узлов

Таким образом, разработан комплексный подход к выявлению доступных IPv6-узлов, сочетающий анализ политик распределения адресного пространства, пассивный сбор адресов через NTP и BitTorrent, поиск по доменным именам, ML-генерацию адресов и прицельное активное сканирование. Предложенное решение позволяет сократить пространство перебора до объёмов, практически достижимых современными инструментами интернет-сканирования.

Жолус А.Р., Чуйков Р.Ю., Кочетков И.В.
Военно-космическая академия имени А.Ф. Можайского, Санкт-Петербург

СИСТЕМА МОНИТОРИНГА ПЕРЕДВИЖЕНИЯ ОТСЛЕЖИВАЕМЫХ ОБЪЕКТОВ С ОТОБРАЖЕНИЕМ НА КАРТЕ

В настоящее время отсутствие или нестабильная работа сотовой связи частое явление. Существует потребность в надёжном мониторинге передвижения людей, транспортных средств или ценных грузов, например, в лесных массивах, горных местностях, удалённых сельскохозяйственных угодьях, а также в районах, подверженных аварийным отключениям инфраструктуры. Представленные на сегодняшний момент трекеры для решения данных задач не подходят, так как полностью зависят от покрытия GSM или наличия интернета, что делает их бесполезными в автономных условиях. Спутниковые трекеры, в свою очередь, характеризуются очень высокой стоимостью как самого оборудования, так и абонентского обслуживания. Кроме того, у подавляющего большинства аналогов низкое время автономной работы, что затрудняет их использование в длительных экспедициях, спасательных работах или для мониторинга на протяжении нескольких суток. В качестве решения возможна система устройств с самоорганизующуюся mesh-сетью, использующая технологию LoRa [1], благодаря которой достигается низкое энергопотребление. Трекер не требует наличия сотовой связи или интернета. Данная разработка представлена на рисунке 1.

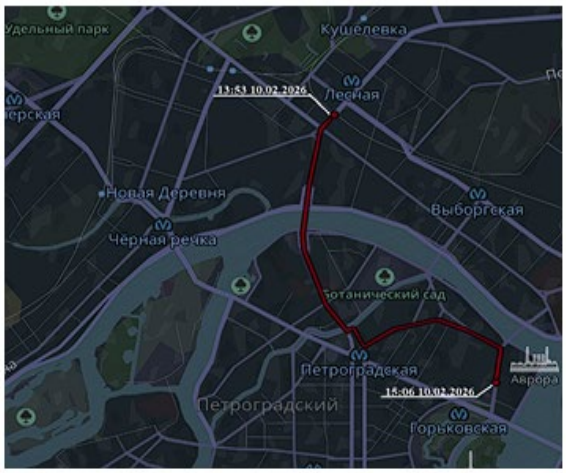
Рисунок 1 – Внешний вид разработанного трекера

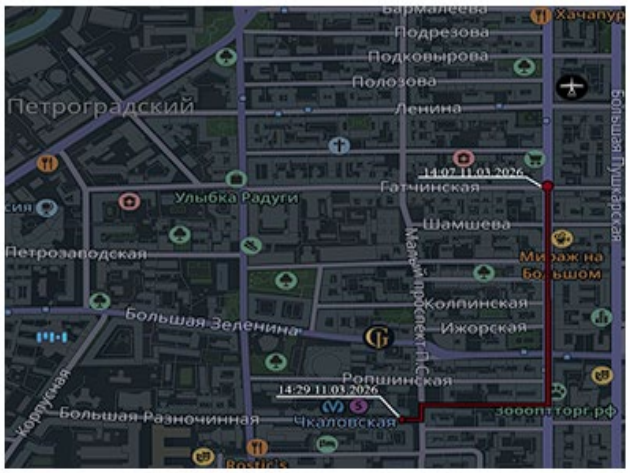
Особенность работы данного устройства заключается в том, что GPS-модуль при необходимости получает от спутников текущие координаты. Эти данные передаются на плату Arduino Uno, которая обрабатывает и упаковывает их в компактный пакет определённой структуры и через SPI-интерфейс передаёт на LoRa-модуль [2]. После осуществляется радиопередача пакета на приёмное устройство, которым может быть другой аналогичный трекер в режиме ретранслятора или специализированный шлюз, подключённый к серверу. Приёмник дешифрирует полученный пакет, извлекает координаты и отображает их. Ключевые особенности - низкое энергопотребление, полная автономность от внешней сети, возможность масштабирования путём добавления новых модулей, которые автоматически находят друг друга. Основные тактико-технические характеристики представлены в таблице 1, а отдельные экранные формы на рисунке 1.

Таблица 1 – Тактико-технические характеристики разработанного устройства

Габаритные размеры (мм)	75 × 65 × 25
Частота обновления координат	Каждую секунду
Энергопотребление в режиме активной передачи данных	50 мА
Максимальная дальность связи на открытой местности	8 км
Точность GPS	3 метра

Питание	5 В (USB/аккумулятор)
Диапазон частот LoRa	433/868 МГц





Для выявления достоинств был проведен анализ с аналогами [3]. Сравнительная характеристика представлена в таблице 2.

Таблица 2 – Сравнительный анализ разработанного устройства с аналогами

Устройства / Характеристики	StarLine M13 ECO	Garmin DriveTrack 72	Petsee 4G	Разработанное устройство
Габаритные размеры (мм)	75 × 33 × 13	95× 60 × 30	⌀48 × 18	75 × 65 × 25
Точность GPS	5–15 м	10 м	15 м	3 м
Питание	90 В	12/24 В	10 В	5 В
Зависимость от сети интернета	Да (через SIM)	Нет (прямой радиоканал)	Да (через SIM)	Нет
Максимальная дальность связи	Зависит от сотовой вышки (5-10 км)	До 8 км (радиоканал)	Зависит от сотовой вышки (5-10 км)	8 км
Примерная стоимость	15000 руб.	25 000 – 35 000 руб.	10000 руб.	8000 руб.

Проведенный анализ показал, что у рассматриваемой системы есть несколько перспективных направлений развития. Первое – это миниатюризация. Вместо Arduino Uno планируется переход на более компактную и энергоэффективную ESP32, а в последствии разработка собственной печатной платы. Второе направление – защита передачи путём внедрения шифрования LoRa-пакетов для предотвращения перехвата.

Разработанная система мониторинга позволяет эффективно отслеживать передвижение объектов в условиях полного отсутствия сотовой связи. Низкое энергопотребление, дальность связи до восьми километров и возможность построения mesh-сети с ретрансляцией сигнала делают данную систему перспективным решением для использования в лесной и горной местности, при проведении поисково-спасательных работ и при ликвидации чрезвычайных ситуаций.

Список литературы:
1. LoRaWAN 1.0.4 Specification: техническая документация / LoRa Alliance. 2020.
2. ATmega328P Microcontroller: техническая документация / Microchip Technology Inc. 2018.
3. NEO–6M / NEO–8M: руководство по работе с GPS-модулями / U-blox, Inc. 2019.

Карапетьянц Н., Никитин Т.А.
Национальный исследовательский ядерный университет «МИФИ», Москва

ОБЕСПЕЧЕНИЕ ИНФОРМАЦИОННОЙ БЕЗОПАСНОСТИ КРИПТОВАЛЮТНЫХ ПЛАТЕЖНЫХ СИСТЕМ: СБОР И МНОГОУРОВНЕВАЯ ФИЛЬТРАЦИЯ ДАННЫХ ОБ УГРОЗАХ

В последние годы использование криптовалют в качестве средства платежа стало частью общемировой финансовой системы. Однако прозрачность блокчейна, которая является одним из его преимуществ, создала уникальную проблему для информационной безопасности. Криптовалютный адрес — это не просто набор символов, а важный индикатор компрометации. Рост активности хакерских группировок, использующих программы-вымогатели, развитие теневых миксеров и участившиеся взломы DeFi-протоколов [1] превратили задачу мониторинга цифровых активов из узкоспециализированной работы криминалистов в обязательный элемент корпоративной системы анализа киберугроз (Threat Intelligence). Проблема усугубляется тем, что злоумышленники постоянно совершенствуют методы обхода автоматизированных систем контроля, используя кросс-чейн-мосты и децентрализованные биржи для обфускации и запутывания транзакционной активности. В этих условиях эффективность защиты криптовалютной платежной системы напрямую зависит не только от внутренних политик безопасности, но и от качества, полноты и скорости обновления внешних ресурсов данных об угрозах информационной безопасности. Без интеграции актуальных данных об угрозах в реальном времени современные организации, использующие криптовалюту в качестве средства платежа, могут быть непреднамеренно вовлечены в незаконную деятельность. Актуальность исследования обусловлена необходимостью формирования унифицированного информационного ландшафта, в котором сведения о специфических блокчейн-угрозах признаются фундаментальным элементом архитектуры информационной безопасности криптовалютных платежных систем, сопоставимым по значимости с базами традиционных индикаторов компрометации. Систематизация источников данных об угрозах является частью архитектуры системы непрерывного мониторинга инфраструктуры криптовалютных платежных систем. Их можно разделить на следующие три категории:

- открытые: публичные базы, отчеты компаний по информационной безопасности, ветки форумов и т.д;
- коммерческие поставщики: Chainalysis, Elliptic, TRM Labs, Crystal Blockchain и т.д;
- источники сводок данных об угрозах, содержащие не только адреса, но и связанные с ними IP, домены фишинговых сайтов и хеши вредоносного программного обеспечения.

К первой категории источников данных относятся специализированные базы данных (например, BitcoinAbuse, Scam Search), а также публичные реестры меток в блокчейн-обозревателях (Etherscan, Blockchain.com). Ключевым преимуществом этих ресурсов является высокая скорость наполнения за счет сообщества, однако они характеризуются переменным уровнем верификации. Для специалистов информационной безопасности они служат первичным сигналом, требующим обязательной кросс-проверки для минимизации риска ложноположительных срабатываний. Также в данную категорию входят источники

данных, ориентированные на закрытые сегменты сети и профильные форумы. Сбор данных в этой области позволяет идентифицировать реквизиты злоумышленников на этапе подготовки атак или в процессе продажи украденных данных или средств. Мониторинг подобных ресурсов обеспечивает наиболее раннее оповещение об угрозах, что дает возможность платежным системам блокировать потенциально опасные финансовые потоки еще до их фактического появления в распределенном реестре.

Вторая категория источников данных представлена коммерческими аналитическими платформами, такими как Chainalysis, Elliptic или TRM Labs [2]. Данные ресурсы аккумулируют информацию путем глубокого эвристического анализа и кластеризации транзакций, что позволяет идентифицировать принадлежность адресов конкретным субъектам — от легитимных бирж до даркнет-маркетплейсов. Использование этих источников критически важно для выполнения требований криптокомплаенса, так как они предоставляют не просто статические списки, а динамический расчет риска каждой отдельной операции или контрагента.

Третью категорию формируют специализированные потоки данных об угрозах, адаптированные для интеграции в классические системы защиты SIEM. В отличие от финансово-ориентированных сервисов, данные источники фокусируются на технической взаимосвязи криптоактивов с другими индикаторами компрометации [3]. Они связывают адреса кошельков с IP-адресами управляющих серверов ботнетов, доменами фишинговых сайтов и контрольными суммами вредоносного ПО, позволяя рассматривать криптовалютную транзакцию как элемент комплексной кибератаки.

Сбор данных об угрозах из вышеперечисленных источников основывается на многоэтапной агрегации индикаторов, включающей автоматизированное извлечение сведений через API-интерфейсы, эвристический анализ транзакционных паттернов и экспертную верификацию инцидентов. Однако критическим дестабилизирующим фактором в этом процессе выступает риск «отравления» данных, который реализуется как через преднамеренную дезинформацию в публичных ресурсах, так и посредством проведения «пылевых атак» [4]. Подобные манипуляции направлены на искусственное связывание легитимных адресов с кошельками злоумышленников для снижения их репутационного профиля, что создает существенные помехи в работе систем мониторинга и требует применения продвинутых алгоритмов кросс-валидации для верификации поступающих сведений.

Синтез данных из различных категорий источников позволяет сформировать многоуровневую систему фильтрации, минимизирующую влияние недостоверной информации и намеренных манипуляций. Практическая значимость данного подхода заключается в возможности построения адаптивной скоринговой модели, в которой весовой коэффициент каждого источника определяется его релевантностью, исторической точностью и подтвержденным уровнем достоверности. Внедрение такой иерархической структуры оценки обеспечивает высокую точность атрибуции активов и идентификации сценариев атак, что в конечном итоге существенно повышает общую резистентность к угрозам информационной безопасности в инфраструктуре криптовалютных платежных систем.

Список литературы:

arpentier-Desjardins C. et al. Mapping the DeFi crime landscape: An evidence-based picture //Journal of Cybersecurity. – 2025. – Т. 11. – №. 1. – С. tyae029.

2. Aidoo S. et al. Cryptocurrency and Sanctions Evasion: Stablecoins, Mixers, and the Role of Blockchain Forensics. – 2024.
3. Monterrubio S. M. M., Solís J. F., García J. A. R. THREATMOSAIC: collection, curation and enrichment of indicators of compromise (IOCs) //International Journal of Combinatorial Optimization Problems and Informatics. – 2025. – Т. 16. – №. 3. – С. 441.

4. Prince D., Blackstone J. DFRA for Web3 Security: Leveraging Federated Learning and Decentralized Storage //2025 IEEE International Conference on Decentralized Applications and Infrastructures (DAPPS). – IEEE, 2025. – С. 121-126.

Комаров Т.И., Жуков И.Ю., Половнева Ю.А., Чепик Н.А., Хисамутдинов М.А.
Национальный исследовательский ядерный университет «МИФИ», Москва

СИСТЕМНЫЙ МЕНЕДЖЕР – ЭВОЛЮЦИОННОЕ РАЗВИТИЕ ПАКЕТНОГО МЕНЕДЖЕРА

Пакетным менеджером принято называть системное программное обеспечение (ПО), которое решает основные задачи, связанные с управлением ПО – установка, удаление, обновление, настройка (чаще всего – ограниченная).

Классические пакетные менеджеры (например, rpm в связке с dnf или dpkg совместно с apt) имеют ряд существенных проблем, например, невозможность установки нескольких версий одного и того же пакета или возможность влияния установленных пакетов друг на друга [1]. Лишь некоторые из проблем постепенно решаются в процессе развития данных пакетных менеджеров, либо при разработке их более современных аналогов.

Комплексное решение указанных проблем возможно лишь в пакетных менеджерах, построенных на иных принципах, например, Nix [2] и Guix [3]. Их фундаментальным отличием от классических систем управления пакетами является отказ от FHS [4] (Filesystem Hierarchy Standard, определяющий расположение основных файлов и каталогов в Unix-подобных ОС и, в частности, дистрибутивах ОС на базе ядра Linux) в пользу обособленного хранилища пакетов, в котором размещаются пакеты в каталогах вида `<HASH>-<NAME>-<VER>`, где `<HASH>` – результат работы криптографической хеш-функции, применённой ко всем входным данным пакета, `<NAME>` – имя пакета, `<VER>` – версия пакета [2].

Подобное решение позволяет использовать функциональный подход также для конфигураций системы. При этом система может быть декларативно описана, а пакетный менеджер получает возможность практически целиком её воспроизвести (кроме паролей и пользовательских данных). Пакетный менеджер, реализующий данные принципы, становится системным менеджером – единым центром управления и конфигурации операционных систем (ОС). Появление подобного инструмента позволит найти решение для многих актуальных проблем в области информационной безопасности, характерных для современных ОС.

В апреле-мае 2026 года в различных подсистемах/драйверах ядра Linux были обнаружены уязвимости Copy Fail (CVE-2026-31431), Dirty Frag (CVE-2026-43284 и CVE-2026-43500), Fragnesia (CVE-2026-46300), DirtyDecrypt (CVE-2026-31635), PinTheft (CVE-2026-43494) и GRO Frag (CVE-идентификатор не был присвоен на момент подготовки тезисов), принцип которых заключается в возможности перезаписи злоумышленником данных в страничном кэше. Данные уязвимости позволяют непривилегированному пользователю получить полномочия суперпользователя, для большей части из них были опубликованы эксплойты [5].

Очередное обнаружение нового семейства критических уязвимостей в ядре Linux напоминает о необходимости продолжения дискуссии по актуальным проблемам в области ИБ, часть из которых представлена в таблице №1.

Таблица №1

Актуальные проблемы в области ИБ, характерные для современных Unix-подобных ОС

Описание проблемы	**Возможные варианты решения проблемы**	**Применимость системного менеджера для решения проблемы**
Современные Unix-подобные ОС чаще всего используют монолитные (модульные) ядра, что существенно снижает их надёжность и безопасность.	Микроядерные ОС за несколько десятилетий своего развития доказали, что могут существенным образом поднять уровень безопасности, при этом обеспечивая приемлемый уровень производительности [6].	Микроядерные ОС требуют намного более сложной настройки с точки зрения описания компонентов и возможностей по их взаимодействию между собой, для чего используются специальные инструменты [7]. Использование системного менеджера позволило бы существенно упростить данный процесс.
Современные Unix-подобные ОС предлагают различные механизмы безопасности, но не все из них повсеместно применяются на практике.	Повсеместное внедрение систем мандатного контроля доступа (например, SELinux, AppArmor или МРД в Astra Linux) и их корректная настройка могут существенным образом повысить уровень безопасности (например, системы с корректно настроенными политиками SELinux оказались защищёнными от уязвимостей семейства CopyFail).	Системный менеджер может предложить удобные и гибкие абстракции, необходимые для настройки систем мандатного контроля доступа.
Модель распространения и обновления современных Unix-подобных ОС не позволяет быстро реагировать на проблемы в области безопасности.	Необходим надёжный и безопасный механизм обновлений, обеспечивающий повсеместное внедрение обновлений на практике.	Функциональный пакетный менеджер позволяет обеспечить надёжный и быстрый процесс обновления, делающий известный принцип «работает – не трогай» плохим аргументом для отказа от регулярной установки обновлений.

Следовательно, разработка российского функционального пакетного менеджера с функциями системного менеджера является важной и актуальной задачей, которую необходимо решать, т.к. наблюдается тенденция к росту количества обнаруживаемых в ПО уязвимостей, в т.ч. в связи с активным развитием искусственного интеллекта.

Список литературы:

1. Состояние и перспективы развития пакетных менеджеров / Т. И. Комаров, И. Ю. Жуков, Т. А. Никитин [и др.] // Методы и технические средства обеспечения безопасности информации. – 2024. – № 33. – С. 118-120. – EDN AQMYTK.
2. Dolstra E. The Purely Functional Software Deployment Model [Электронный ресурс] // URL: https://edolstra.github.io/pubs/phd-thesis.pdf (дата обращения: 25.05.2026).

3. Courtès L. Functional Package Management with Guix [Электронный ресурс] // URL: https://arxiv.org/pdf/1305.4584 (дата обращения: 25.05.2026).
4. Filesystem Hierarchy Standard [Электронный ресурс] // URL: https://specifications.freedesktop.org/fhs/latest (дата обращения: 25.05.2026).
5. nformation about the Copy Fail vulnerability, which allows attackers to gain root access on virtually any modern Linux distribution [Электронный ресурс] // https://securelist.com/tr/copyfail-root-linux/119634/ (дата обращения: 25.05.2026).
6. From L3 to seL4: What have we learnt in 20 years of L4 microkernels? / K. Elphinstone, G. Heiser // SOSP 2013 – Proceedings of the 24th ACM Symposium on Operating Systems Principles. – 2013. С. 133-150.
7. CAmkES: A component model for secure microkernel-based embedded systems / I. Kuz, Y. Liu, I. Gorton, G. Heiser // Journal of Systems and Software. – 2007. Volume 80, Issue 5. – С. 687-699.

Коркин И.Ю.
АО Positive Technologies, Москва

МЕТОД УПРАВЛЕНИЯ ПРАВАМИ ДОСТУПА К ОБЪЕКТАМ РАБОЧЕГО СТОЛА WINDOWS ДЛЯ ЗАЩИТЫ ПОЛЬЗОВАТЕЛЬСКОГО ВВОДА

Актуальность и проблема. Актуальность темы обусловлена противоречием: с одной стороны, сохраняется рост атак вредоносного ПО (ВПО) с функциями перехвата клавиатурного ввода, с другой – существующие средства защиты не в полной мере соответствуют развитию современных угроз [1]. Приоритетной целью таких атак является парольно-адресная информация, вводимая пользователем: мастер-пароли менеджеров паролей, пароли к защищённым документам и иная текстовая информация. Особую опасность представляют атаки класса Man-At-The-End (MATE), при которых нарушитель обладает легитимным доступом к системе и возможностью запуска стороннего ПО без эксплуатации уязвимостей [2].

В работе рассматриваются четыре техники перехвата ввода, использующие функции WinAPI: SetWindowsHookEx, GetAsyncKeyState, механизмы Raw Input и DirectInput [3].

Недостаточность существующих решений. Существующие подходы к защите от перехвата клавиатурного ввода можно разделить на три группы: зашумление ввода, обнаружение программ перехвата и перенос защищаемого приложения в изолированную среду. Первые два подхода имеют существенные ограничения: зашумление может быть заблокировано штатными механизмами ОС, а обнаружение кейлоггеров затрудняется при использовании нарушителем обфускации, отложенного запуска и других техник противодействия.

Наиболее перспективным представляется подход на основе штатной технологии Windows Desktop, используемой в Winlogon, KeePass2 и VeraCrypt. Его суть состоит в переносе защищаемого приложения с рабочего стола по умолчанию на отдельно созданный рабочий стол. В этом случае кейлоггеры пользовательского режима, работающие только в контексте исходного рабочего стола, не получают сведений о нажатых клавишах, поскольку Windows обеспечивает изоляцию системных очередей сообщений для каждого рабочего стола.

Однако штатное использование Windows Desktop само по себе не исключает несанкционированное подключение нарушителя к защищаемому рабочему столу [4]. Причина состоит в том, что рабочий стол Windows является именованным объектом, связанным с оконной станцией, дескриптором безопасности и набором прав доступа. Нарушитель даже с привилегиями пользователя может перечислить рабочие столы,

получить имя целевого объекта, открыть его по имени либо запустить процесс в его среде. После такого подключения техники перехвата клавиатурного ввода вновь становятся применимыми и для случая изолированного стола.

Предлагаемое решение. В работе предлагается метод управления правами доступа к объектам рабочего стола Windows. Метод основан на создании отдельного рабочего стола, управлении доступностью его имени, ограничении права перечисления рабочих столов на уровне оконной станции и изменении дескриптора безопасности созданного рабочего стола.

Символьное имя защищённого стола генерируется как псевдослучайная последовательность. Рабочий стол создаётся с дескриптором безопасности, заданным в виде SDDL-описания «D:P», в результате имя стола не возвращается при перечислении, а попытка открыть объект по имени с использованием штатных функций ОС завершается неуспешно. При этом управляющий процесс сохраняет возможность переключения пользователя между столами.

Для запуска пользовательского приложения в среде защищённого стола выполняется кратковременное расширение прав доступа, и включает в себя следующие этапы. 1. Отзыв прав доступа «winsta_enumdesktops» у объекта оконной станции для запрета получения списка рабочих столов. 2. Расширение прав доступа к защищённому столу для SID текущего пользователя, SDDL-описание: «D:(A;;GA;;;SID_u)». 3. Запуск приложения в защищённой среде и ожидание завершения его инициализации. 4. Восстановление значения дескриптора безопасности, SDDL-описание «D:P». 5. Восстановление прав доступа «winsta_enumdesktops» у объекта оконной станции. Таким образом, момент расширения прав не сопровождается раскрытием имени рабочего стола.

Формализация метода. Метод формализован на уровне состояний системы. Состояние определяется двумя признаками: e(t) – возможностью нарушителя перечислять рабочие столы через право «winsta_enumdesktops», и g(t) – возможностью использовать символьное имя защищённого рабочего стола для открытия объекта или запуска процесса в его контексте. Опасным считается состояние, в котором одновременно выполняются условия e(t)=1 и g(t)=1, то есть нарушитель способен узнать имя защищённого рабочего стола и использовать это имя для подключения. Динамика переходов описывается в терминах теории графов: вершины соответствуют состояниям доступа, а рёбра – операциям изменения прав оконной станции, изменения DACL рабочего стола и запуска приложения. Требование безопасности формулируется как недостижимость опасного состояния при штатной работе метода. Вероятность подбора имени стола пренебрежимо мала.

Экспериментальная проверка. Подход реализован в исследовательском прототипе на языке C++. Были проведены экспериментальные испытания с опытным прототипом нарушителя, реализующим перехват клавиатурного ввода и несанкционированное подключение к рабочим столам пользователя. Сравнивались существующие механизмы защиты и предлагаемый подход (см. табл. 1.). Знак «+» означает успешную регистрацию нажатий клавиш соответствующей техникой на данном рабочем столе, знак «–» – отсутствие возможности перехвата.

Таблица 1 — Сравнение механизмов защиты пользовательского ввода

Сценарий тестирования	Результаты тестирования техники перехвата на различных рабочих столах		
	KeePass Secure Desktop	Winlogon Secure Desktop	**Предлагаемый подход**
Кейлоггер в режиме подключения к рабочим столам (привилегии пользователя)	+	–	**–**

Кейлоггер в режиме подключения к рабочим столам (системные привилегии «nt authority\system»)	–	+	**–**
Возможность запуска произвольного пользовательского приложения в защищённой среде	нет	нет	**да**

Заключение. Предлагаемый подход, в отличие от существующих решений, блокирует штатный сценарий подключения кейлоггера к защищённому рабочему столу, в том числе при запуске нарушителя с системными привилегиями в рамках принятой модели угроз. Кроме того, подход обеспечивает запуск произвольных пользовательских приложений в защищённой среде.

Список литературы:

1. Fortinet. (2024). Deep analysis of Snake Keylogger's new variant. Fortinet Blog. Retrieved April 4, 2026, URL: https://www.fortinet.com/blog/threat-research/deep-analysis-of-snake-keylogger-new-variant
2. Ianni, M., & Schrittwieser, S. (2025). Special issue on offensive and defensive techniques in the context of Man-At-The-End (MATE) attacks. ACM Transactions on Privacy and Security, 28(2), Article 31. DOI:https://doi.org/10.1145/3768371
3. Sajid M. S. I., Ahmed S., Sosnoski R. Secure Development of a Hooking-Based Deception Framework Against Keylogging Techniques // 2025 IEEE Secure Development Conference (SecDev). IEEE, 2025. P. 82–91. DOI: 10.1109/SecDev66745.2025.00019
4. Denkiewicz A. Master Password can be keylogged from "Secure desktop" — URL: https://github.com/adenkiewicz/KeepassKeylogger (дата обращения: 10.10.2025).

Крюков Р.О., Шаповалов Д.Г.

Военно-космическая академия им. А.Ф. Можайского, Санкт-Петербург

ПОДХОД К АВТОМАТИЗИРОВАННОМУ АУДИТУ И КОМПЛАЕНС-КОНТРОЛЮ KUBERNETES-КЛАСТЕРОВ С ИСПОЛЬЗОВАНИЕМ ANSIBLE

Современные корпоративные и государственные информационные системы всё чаще строятся на базе контейнеризованной инфраструктуры и Kubernetes-кластеров. Несмотря на гибкость и широкое распространение, Kubernetes включает множество взаимосвязанных компонентов, ошибки в настройке которых могут повлечь за собой избыточные права доступа, нарушение контейнерной изоляции, расширение поверхности атаки и снижение отказоустойчивости ИС.

Одной из главных проблем в эксплуатации Kubernetes является отклонение между фактическим состоянием кластера и эталонной безопасной конфигурацией. На практике это проявляется в небезопасных настройках API Server и kubelet; отсутствии аудита событий безопасности; избыточных правах RBAC; запуске привилегированных контейнеров; использовании hostPath, hostNetwork, hostPID, hostIPC; отсутствии базовых NetworkPolicy. Ручной контроль таких параметров требует значительных ресурсов, зависит от квалификации администратора и не обеспечивает систематическую доказательную базу соблюдения требований безопасности.

Актуальность автоматизации подтверждается широким кругом нормативных, методических и отраслевых требований к защите ИС. В качестве практического технического эталона для Kubernetes может служить CIS Kubernetes Benchmark, который

позволяет формализовать проверяемые требования к элементам кластера и связать их с задачами аудита и комплаенс-контроля.

Для решения поставленной задачи разработана методика автоматизированного аудита и комплаенс-контроля Kubernetes-кластера на базе Ansible. Архитектура решения построена по принципам Infrastructure as Code и Compliance as Code. Ansible является ядром подсистемы: проводит аудит control plane и worker nodes, сверяет параметры kubelet, права доступа к критичным файлам, конфигурации RBAC и Service Accounts, наличие сетевых политик, запускает kube-bench и агрегирует результаты Kyverno. kube-bench используется как независимый инструмент проверки соответствия CIS Kubernetes Benchmark, а Kyverno – как policy-as-code механизм контроля Kubernetes-ресурсов внутри кластера.

Согласно архитектуре подхода, пользователь через Web UI выбирает режим работы, перечень проверяемых доменов и запускает аудит или ремедиацию. API Semaphore используется как промежуточный слой для запуска Ansible-плейбуков из веб-интерфейса. После выполнения проверок результаты Ansible, kube-bench и Kyverno передаются в Collector/Normalizer, где приводятся к единому формату: контроль, домен безопасности, статус, критичность, ожидаемое и фактическое значение и источник проверки. В Web UI отображаются общий уровень соответствия, перечень нарушений, их приоритет, история запусков и состояние до/после выполнения комплаенс-контроля.

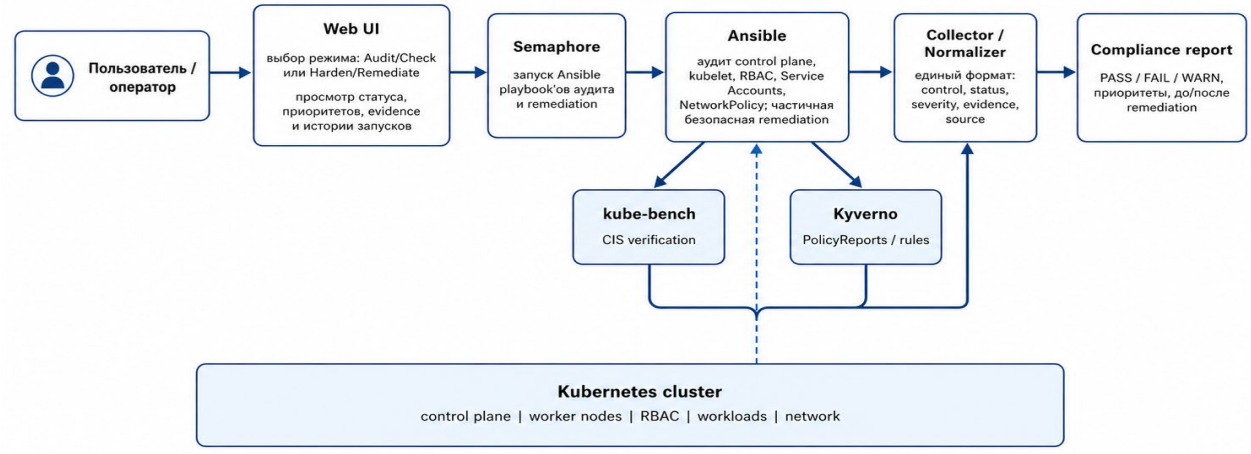


Рисунок 1. Архитектура подхода

Предлагаемый подход обладает следующими ключевыми преимуществами по сравнению с существующими инструментами:

1) Процесс автоматизирован и снижает ошибки ручной проверки. Реализуются два режима работы:
   - Audit/Check – аудит текущего состояния без внесения изменений, с формированием отчёта;
   - Harden/Remediate – применение безопасно автоматизируемых мер устранения нарушений с последующей повторной проверкой. Ansible содержит роли и плейбуки, описывающие желаемое состояние кластера по выбранным доменам безопасности.
2) Масштабируемость. Подход применим не к одному узлу, а ко всем узлам кластера.
3) Повторяемость и идемпотентность. Повторный запуск Ansible не нарушает уже корректно настроенные параметры, а приводит систему к заданному эталонному состоянию.
4) Прозрачность и документируемость. Результаты проверок до и после remediation сохраняются в виде отчета, что позволяет подтвердить выявленные нарушения, выполненные изменения и итоговый статус соответствия.
5) Приоритизация и практическая применимость. В отличие от решений класса Wazuh, которые в большей степени ориентированы на агрегацию и визуализацию состояния

безопасности, предлагаемая подсистема позволяет определить, какие несоответствия необходимо закрывать в первую очередь.

Таким образом, предлагаемый автоматизированный подход позволяет выстроить централизованную, гибкую и проверяемую систему аудита и комплаенс-контроля Kubernetes-кластера, превращая безопасность из разовой ручной проверки в управляемый, измеримый и непрерывно контролируемый процесс.

Список литературы

1. ФСТЭК России. Методический документ «Мероприятия и меры по защите информации, содержащейся в информационных системах». Раздел 4.5 «Защита технологий контейнерных сред и их оркестрации». — 2026.
2. Михаил Климарев, Антон Шумаков. Ansible для DevOps. Автоматизация управления конфигурациями и развёртывания. – СПб.: Питер, 2022.
3. Aqua Security. kube-bench: CIS Kubernetes Benchmark checks [Электронный ресурс]. – URL: https://github.com/aquasecurity/kube-bench (дата обращения: 01.05.2026г.).
4. Center for Internet Security. CIS Kubernetes Benchmark v.1.11.1. — CIS, 2023.
5. Kubernetes Documentation. Kubernetes Components; Pod Security Standards; RBAC Authorization.
6. Kyverno. Kyverno Documentation: Automation and Configuration Management.

Кузнецов Н.Е., Карасев В.Т., Резанов Д.И.

Военно-космическая академия имени А.Ф. Можайского, Санкт-Петербург

ГИБРИДНАЯ СХЕМА ШИФРОВАНИЯ ИЗОБРАЖЕНИЙ С ИСПОЛЬЗОВАНИЕМ КАРТЫ ХЕНОНА С ДИНАМИЧЕСКОЙ ГЕНЕРАЦИЕЙ КЛЮЧЕЙ

Безопасность передачи изображений является одним из ключевых аспектов современной информационной безопасности. Одной из актуальных задач в данной области является обеспечение устойчивости алгоритмов шифрования к различным видам атак при сохранении эффективности обработки данных. Особую сложность представляет защита изображений, поскольку они характеризуются высокой избыточностью и сильной корреляцией между соседними пикселями [1].

Шифрование изображений направлено на преобразование визуальной информации в форму, недоступную для несанкционированного доступа. Обычно для этих целей применяются как классические криптографические методы, так и алгоритмы, основанные на хаотических системах [2]. Однако даже при использовании современных подходов остаются возможные уязвимости, связанные с предсказуемостью ключей и недостаточной чувствительностью к входным данным.

В настоящее время для защиты изображений от несанкционированного доступа широко используется карта Хенона, представляющая собой дискретную динамическую систему, обладающую свойствами хаотичности и высокой чувствительности к начальным условиям [3]. Математически преобразование на основе карты Хенона описывается следующей системой уравнений:

$$\begin{cases} x_{n+1} = 1 - ax_n^2 + y_n, \\ y_{n+1} = bx_n, \end{cases}$$

где x_n, y_n – переменные состояния системы на n-м шаге; a, b – управляющие параметры, определяющие режим динамики; x_0, y_0 – начальные условия, малейшее изменение которых приводит к экспоненциально расходящимся траекториям. На основе представленных преобразований формируются ключевые последовательности, применяемые в процессах нелинейной подстановки и диффузии. Тем не менее, в классических схемах шифрования

генерация ключей осуществляется на основе фиксированных параметров, что может снижать криптостойкость алгоритма [3].

В данной работе предлагается использование динамической генерации ключей на основе карты Хенона, при которой параметры системы зависят от характеристик исходного изображения. Это позволит повысить уровень безопасности за счет увеличения пространства ключей и устранения повторяемости ключевых последовательностей.

Пример применения гибридной схемы шифрования

Предположим, у нас есть алгоритм шифрования изображения, реализованный следующим образом [1]:

1) исходное изображение подвергается перестановке пикселей;
2) изображение разбивается на блоки фиксированного размера;
3) генерируется ключевая матрица с использованием карты Хенона;
4) выполняется операция XOR между блоками изображения и ключом;
5) применяется диффузия для распространения изменений по всему изображению.

Данный алгоритм реализует классический подход к хаотическому шифрованию. Однако существует ограничение – параметры карты Хенона остаются неизменными, что делает систему уязвимой к анализу при многократном использовании [3].

В предлагаемом подходе генерация ключей модифицируется следующим образом:

1) вычисляется хэш исходного изображения;
2) полученное значение используется для определения начальных условий карты Хенона;
3) параметры a и b динамически изменяются в зависимости от полученого хэша;
4) для каждого изображения формируется уникальная ключевая последовательность.

На рисунке 1 представлена схема работы алгоритма шифрования изображения с динамическим изменением параметров карты Хенона.

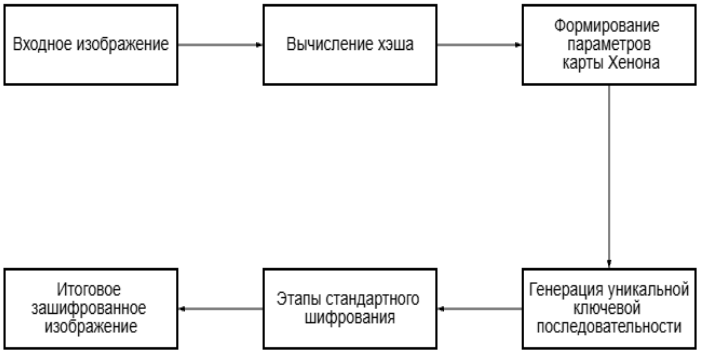


Рисунок 1. – Схема основных этапов шифрования изображения с динамическим изменением параметров карты Хенона

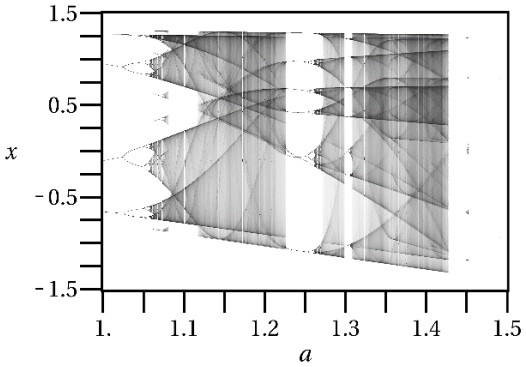


Рисунок 2. – Диаграмма изменения параметров карты Хенона в зависимости от изменения параметра a

Порядок работы алгоритма следующий.

1. Входное изображение анализируется и вычисляется его хэш.
2. На основе хэша формируются параметры карты Хенона.
3. Генерируется уникальная ключевая последовательность
4. Выполняются стандартные этапы шифрования (нелинейная подстановка, XOR, диффузия).

В результате работы алгоритма формируется зашифрованное изображение.

Использование динамической генерации ключей позволяет устранить повторяемость ключевых последовательностей, характерную для классических схем [3]. Это существенно усложняет проведение криптоанализа, поскольку даже незначительное изменение входного изображения приводит к изменению всей ключевой структуры.

На рисунке 2 представлена бифуркационная диаграмма карты Хенона, иллюстрирующая установившиеся значения переменной *x* в зависимости от изменения управляющего параметра *a* в диапазоне от 1,1 до 1,5 [4]. Диаграмма демонстрирует чередование областей периодического и хаотического поведения, что подтверждает высокую чувствительность системы к выбору параметра *a.* Задание параметра *a* на основе хеша изображения переводит ключевую последовательность в непредсказуемый для злоумышленника режим и делает атаки, основанные на анализе ключей, практически неосуществимыми.

Таким образом, внедрение динамической генерации ключей в гибридую схему шифрования изображений на основе карты Хенона позволяет повысить криптографическую стойкость алгоритма. Направлением дальнейшего развития предложенной схемы гибридного шифрования является оптимизация алгоритмов для работы в реальном времени.

Список литературы:

1. Zhang B., Liu L. Chaos-Based Image Encryption: Application, and Challtnges // MDPI. URL: https://www.mdpi.com (дата обращения: 21.04.2026).
2. Kocarev L. Chaos-Based Cryptography // IEEE Circuits and Systems Magazine. - 2001. - Vol. 1, No. 3. - P. 6-21. - RUL: https://www.researchgate.net (дата обращения: 21.04.2026).
3. Inam S. [и др.] A Novel Hybrid Image Encryption Scheme Using Henon Map for Secure Image Commenication // IET Resarch Journals. Wiliy. URL: https://ietresearch.onlinelibrary.wiley.com (дата обращения: 21.04.2026).
4. Lee S. Unlocking the Hénon Map Secrets: A Comprehensive Guide to Dynamical Systems and Chaos Theory [Электронный ресурс] // Number Analytics. 2025. 28 May. URL: https://www.numberanalytics.com/blog/ultimate-guide-to-henon-map (дата обращения: 04.05.2026).

Макеева А.А.

Санкт-Петербургский политехнический университет Петра Великого, Санкт-Петербург

МЕТОД ВЫЯВЛЕНИЯ ТАРГЕТИРОВАННЫХ АТАК СОЦИАЛЬНОЙ ИНЖЕНЕРИИ В ДИАЛОГАХ НА ОСНОВЕ КВАЗИ-ИДЕНТИФИКАТОРОВ ПОЛЬЗОВАТЕЛЯ

Атаки социальной инженерии в последнее время характеризуются высокой степенью персонализации, которая достигается за счет предварительного сбора сведений о потенциальной жертве [1]. Увеличение объема доступной злоумышленнику информации привело к возможности адаптировать сценарии воздействия под конкретного пользователя. Большинство исследований на данный момент ориентированы на выявление универсальных признаков фишинга, таких как лексические особенности текста, характеристики URL-адресов, свойства вложений, а также метаданные электронных сообщений [2, 3]. В ряде работ отмечается, что злоумышленники используют данные о жертве для входа в доверие [4], однако авторы не рассматривают персонализацию как значимый критерий наличия атаки. Кроме того, такие исследования в основном посвящены анализу отдельных сообщений электронной почты [5], тогда как атаки в форме телефонного мошенничества реализуются в виде диалога. В результате возрастает актуальность

разработки методов автоматизированного выявления таргетированных атак социальной инженерии в диалогах.

Под таргетированной атакой социальной инженерии в рамках исследования понимается коммуникационное воздействие, направленное на достижение цели злоумышленника посредством использования сведений о конкретной жертве [6]. В качестве критерия таргетированного воздействия рассматривается невозможность применения сообщения к произвольному пользователю без изменения его содержания. Персонализация (таргетинг) повышает доверие жертвы к злоумышленнику, т.е. увеличивает вероятность выполнения жертвой требуемого действия [4].

Квази-идентификаторы представляют собой признаки, которые не позволяют однозначно идентифицировать пользователя по отдельности, однако могут использоваться для идентификации при совместном рассмотрении [7]. К таким признакам могут относиться возраст, регион проживания, место работы, образование, профессиональная деятельность, интересы, социальные связи и иные сведения, встречающиеся в открытых источниках. Использование квази-идентификаторов позволяет формировать персонализированные сообщения, следовательно, – имитировать доверительное общение и повышать эффективность манипулятивного воздействия. Например, в сообщении «Скотт, это Кристофер Далбридж. Я только что разговаривал по телефону с мистером Биггли, и он очень недоволен. Он говорит, что десять дней назад отправил вам записку с просьбой предоставить нам копии всех ваших исследований проникновения на рынок для анализа. Мы ничего не получили» используются сведения о профессиональной деятельности и рабочем окружении жертвы, что повышает вероятность доверия к сообщению.

Предлагаемый метод выявления таргетированных атак социальной инженерии предполагает анализ диалогов с целью обнаружения признаков персонализации (таргетинга) и признаков склонения жертвы к достижению целей злоумышленника. Метод включает в себя следующие этапы: сбор транскрипции диалога, выявление упоминаний квази-идентификаторов, анализ признаков манипулятивного воздействия и принятие решения о наличии атаки в рассматриваемом контексте. Предполагается, что наличие квази-идентификаторов в сочетании с признаками манипуляции может рассматриваться как индикатор таргетированной атаки социальной инженерии, тогда как использование только признаков манипуляции не позволяет надежно выявлять таргетированные атаки, так как манипулятивные техники могут применяться и в легитимных коммуникациях, например, в служебных уведомлениях или маркетинговых сообщениях.

В качестве реализации предложенного подхода предлагается использование больших языковых моделей для обработки естественного языка и анализа контекста диалога. Были использованы наборы данных, содержащие тексты телефонных разговоров – как мошеннических, так и легитимных. На основе этих данных был составлен перечень квази-идентификаторов, встречающихся в мошеннических диалогах. Затем был создан классификатор, на основе упоминания квази-идентификаторов определяющий, является ли входной текст мошенническим. В рамках валидации предполагается проведение оценки качества выявления атак с использованием метрик классификации. Средство выявления таргетированных атак на основе контекстного анализа упоминания квази-идентификаторов пользователя может быть использовано при разработке интеллектуальных средств мониторинга коммуникаций, например, для создания средства автоматизированного анализа содержания телефонных звонков на предмет мошенничества, обеспечивая также мониторинг распространения персональной информации.

Список литературы:

1. Банк России. Операции, совершенные без добровольного согласия клиентов финансовых организаций. Обзор отчетности за 2024 год [Электронный ресурс]. – URL: https://www.cbr.ru/analytics/ib/operations_survey/2024/ (дата обращения: 19.05.2026).
2. Altwaijry N. et al. Advancing phishing email detection: A comparative study of deep learning models //Sensors. – 2024. – Т. 24. – №. 7. – С. 2077.
3. Evans K. et al. RAIDER: Reinforcement-aided spear phishing detector //International Conference on Network and System Security. – Cham : Springer Nature Switzerland, 2022. – С. 23-50.
4. Park H. et al. Enhanced Voice Phishing Detection Using an LLM-Based Framework for Data Augmentation and Classification //IEEE Access. – 2025.
5. Li Q., Cheng M. Spear-Phishing Detection Method Based on Few-Shot Learning //International Symposium on Advanced Parallel Processing Technologies. – Singapore : Springer Nature Singapore, 2023. – С. 351-371.
6. Advanced social engineering attacks [Электронный ресурс] // Journal of Information Security and Applications. – 2014. – URL: https://www.sciencedirect.com/science/article/pii/S2214212614001343 (дата обращения: 19.05.2026).
7. Сачава П. Д. Защита персональных данных: международные стандарты и национальное регулирование для специальности 6-05-0421-02" Международное право", 6-05-0421-03" Экономическое право". – 2025.

Мангиров П.В., Паращук И.Б., Сабиров Д.Э.
Военная академия связи им. Маршала Советского Союза С.М. Буденного, Санкт-Петербург

АНАЛИЗ ПРЕДПОЧТЕНИЙ В РАМКАХ ПОЛНОМОЧИЙ ВЫБОРА СОВРЕМЕННЫХ СРЕДСТВ МНОГОФАКТОРНОЙ АУТЕНТИФИКАЦИИ КАК ИНСТРУМЕНТА СОЗДАНИЯ БЕЗОПАСНОЙ СРЕДЫ ДАТА-ЦЕНТРОВ ДЛЯ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДАННЫХ

Дискреционные полномочия, зачастую называемые полномочиями и правами выбора должностного лица в рамках анализа, сопоставления достоинств и недостатков специальных изделий и программ, и, самое главное, в рамках принятия решений по обоснованному подбору для проектирования, приобретения (покупки) и построения современных программно-аппаратных инструментов защиты сложных информационно-телекоммуникационных объектов, например, средств и комплексов многофакторной аутентификации (МФА) пользователей современных дата-центров, относятся к трудно формализуемым, и, в большинстве случаев, субъективным ситуациям и задачам, поэтому, требуют новых математически корректных и рациональных инструментов поддержки принятия подобных решений с учетом неопределенных, нечетких предпочтений лиц, вынужденных принимать подобные решения [1, 2].

Речь идет о гарантированной защите от несанкционированного доступа (НСД) и создании безопасной среды в интересах построения и функционирования современных дата-центров (ДЦ), ориентированных, помимо прочего, на выполнение задач по интеллектуальному анализу данных (ИАД). При этом сложности в рамках ИАД, с точки

зрения безопасности, добавляет именно доступ пользователей к ресурсам ДЦ для задач, традиционно решаемых «открытыми» ресурсами: задач обработки и моделирования данных, написания программ на SQL или NoSQL для работы с базами данных, для работы с алгоритмами математической статистики и машинного обучения, включая инструменты обработки «Больших Данных» – общедоступные пакеты Apache Spark, Storm, Samza, Flink или Apache Hadoop, инструменты и библиотеки для визуализации результатов интеллектуального анализа больших объемов данных [3, 4].

Не секрет, что особое место в ряду задач интеллектуального анализа больших объемов данных, решаемых в рамках функционирования современных ДЦ такого класса, отводится вопросам защиты рабочих мест, процедур и результатов такого анализа от НСД, причем одним из наиболее рациональных и не очень затратных подходов при решении подобных вопросов принято считать применение средств и комплексов МФА [5, 6].

Вместе с тем, роль решения задачи обоснованного выбора средств МФА, как инструмента создания безопасной среды дата-центров для интеллектуального анализа данных, неуклонно и перманентно возрастает, особенно в той ее части, которая относится к учету результатов рационального сочетания «цена/качество» в условиях неопределенности, нечеткости исходных данных о приоритетах (предпочтениях, альтернативах) в рамках полномочий такого выбора.

Анализ существующих подходов к выбору средств обеспечения безопасности показывает, что выбор, например, аппаратных и программных средств МФА пользователей критически важных инфраструктур, не в полной мере удовлетворяют требованиям по качеству и объективности подбора недорогих, но производительных, эффективных изделий такого класса. С учетом этого, исключительную значимость, на наш взгляд, приобретает формулировка этапов и учет особенностей в рамках формирования новой методики повышения качества выбора средств МФА, рассматриваемых, как инструмент создания безопасной среды ДЦ для ИАД, с использованием методов теории нечетких множеств, например, методов нечетких отношений предпочтения должностных лиц, вынужденных осуществлять подобный выбор, и методов анализа нечетких альтернатив выбора аппаратных и программных средств многофакторной аутентификации [7].

Анализ и учет нечетких предпочтений и альтернатив в рамках полномочий выбора в интересах последующего приобретения и применения должностными лицами конкретных современных средств МФА, является структурной фазой и важной операционной процедурой поддержки принятия решений, отдельным этапом новой методики повышения качества выбора инструментов аутентификации, в ряду нескольких важных функциональных этапов, таких, как:

задание первичных значений (либо диапазона значений) альтернатив для критериев (показателей, соотношений) выбора типа «цена/качество» для средств МФА, определение нечетких переменных для показателей качества средств аутентификации;

каждый из этих критериев (показателей, соотношений) выбора средств МФА типа «цена/качество» формулируется в виде некоторой переменной, значения этих переменных задаются на конкретном универсальном множестве типов МФА, здесь же формируется вид функции принадлежности данной переменной с учетом нечетких предпочтений и альтернатив в рамках полномочий выбора;

для всех критериев (показателей) типа «цена/качество» составляются матрицы нечетких отношений предпочтения; определяются подмножества недоминирующих альтернатив на нечетком множестве средств МФА по всем выбора нечетким переменным выбора;

определяется множество недоминирующих альтернатив выбора средств МФА как пересечение подмножеств предпочтений.

Использование на практике предложенных этапов в рамках реализации процедур обоснованного выбора современных средств многофакторной аутентификации, осуществление сформулированных практических шагов, основанных на методах нечетких

отношений предпочтения, методах анализа и сравнения альтернатив выбора, позволяет повысить обоснованность принимаемых решений по оптимальному подбору конкретных инструментов аутентификации, и, в конечном итоге, позволяет повысить объективность принимаемых решений по построению безопасной информационно-вычислительной среды для интеллектуального анализа данных.

Список литературы:

1. Пресняков М.В. Дискреционные полномочия и правовая определенность // Власть. 2013. № 5. С. 163-166.
2. Хэммонд Дж., Кини Р., Райффа Г. Правильный выбор. Практическое руководство по принятию взвешенных решений. – М.: МИФ, 2018. 220 с.
3. Hayes G. Six Skills That Will Take Your Data Science Career to the Next Level. [2019]. URL: https://habr.com/ru/companies/plarium/articles/465055/ (дата обращения: 21.03.2026).
4. Барретт А. 10 необходимых навыков для работы с данными. Octoparse. [2022]. URL: https://www.octoparse.com/blog/10-must-have-skills-for-data-mining (дата обращения: 21.03.2026)
5. Карманов А.Г., Галимов Т.А. Средства многофакторной аутентификации в современной инфраструктуре безопасности информационных систем // Info Security. 2012. № 1. С. 94-97.
6. Морозов И.В., Сундуков В.А., Паращук И.Б. Требования к средствам многофакторной аутентификации пользователей в интересах защиты информации в критических инфраструктурах // VIII Межрегиональная научно-практическая конференция «Перспективные направления развития отечественных информационных технологий (ПНРОИТ-2022)». Материалы конференции. – Севастополь: СевГУ, 2022. С. 21-24.
7. Саяркин Л.А., Паращук И.Б., Владимирова Е.С. Этапы и особенности разработки методики повышения качества информационного поиска на ресурсах современных центров обработки данных с использованием нечетких отношений предпочтения и сравнения альтернатив // Информация и космос. 2024. № 1. С. 38-45.

Махорин И.В., Светикова Н.А.

Национальный исследовательский ядерный университет «МИФИ», Москва

МЕТОД ГРАФОВОЙ КОРРЕЛЯЦИИ СОБЫТИЙ С ВЫЯВЛЕНИЕМ ДИНАМИЧЕСКИХ АНОМАЛИЙ В КОРПОРАТИВНОЙ ВЫЧИСЛИТЕЛЬНОЙ СЕТИ

В современных корпоративных вычислительных сетях обнаружение инсайдерских угроз остается сложной задачей информационной безопасности. Инсайдерская активность осуществляется с использованием легитимных учетных записей, что затрудняет ее выявление [1]. Традиционные методы анализа ориентированы на обработку отдельных журналов событий и не учитывают взаимосвязи между действиями пользователей, что снижает их эффективность при обнаружении сложных сценариев атак [2, 6]. Целью работы является разработка метода графовой корреляции событий с выявлением динамических аномалий для обнаружения инсайдерской активности в корпоративной вычислительной сети. Предлагаемый подход основан на использовании динамического графа взаимодействий, в котором вершины соответствуют пользователям, ресурсам и операциям, а ребра отражают факты взаимодействия между ними. Предполагается, что нормальная

деятельность формирует устойчивые паттерны, тогда как инсайдерская активность приводит к появлению нетипичных связей и изменению структуры подграфов.

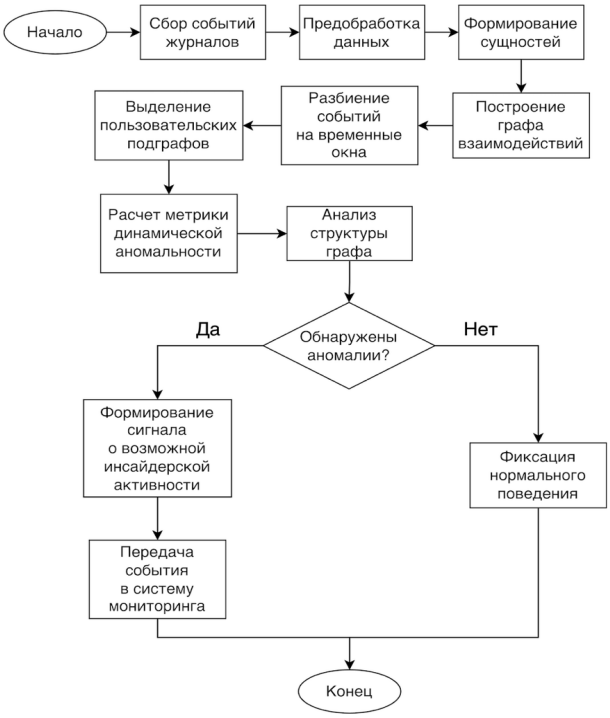


Рисунок 1 – Блок-схема метода графовой корреляции событий с выявлением динамических аномалий

Метод включает следующие этапы: сбор и предварительную обработку событий журналов, формирование сущностей (пользователь, ресурс, действие), построение графа взаимодействий, разбиение на временные окна и формирование последовательности графов, выделение пользовательских подграфов и анализ их структуры. Аномалии выявляются на основе появления новых связей, изменения плотности и формирования нетипичных маршрутов взаимодействий.

Для количественной оценки аномальности введена метрика динамической аномальности:

$$A_t = \alpha \cdot \Delta E_t + \beta \cdot \Delta D_t + \gamma \cdot \Delta P_t \,, \quad (1)$$

где A_t— уровень аномальности в момент времени t;

ΔE_t — изменение числа ребер (новые связи);

ΔD_t — изменение средней степени вершин (плотности);

ΔP_t — изменение количества нетипичных путей;

α, β, γ— весовые коэффициенты.

Метрика рассчитывается относительно базового состояния графа нормальной активности, что позволяет выявлять отклонения от типичных паттернов поведения. Для оценки эффективности метода проведено сравнительное исследование с традиционными подходами.

Таблица 1: Результаты оценки эффективности предложенного метода графовой корреляции событий

Метод анализа	**Точность**	**Полнота**	**Ложные срабатывания**
Анализ отдельных событий	0.71	0.65	0.22

Статический граф	0.79	0.73	0.18
Динамический граф	**0.87**	**0.82**	**0.11**

Результаты показывают, что использование динамического графового представления повышает качество обнаружения инсайдерской активности за счет учета структуры взаимодействий и их изменений. Предложенный метод может быть использован в системах мониторинга информационной безопасности для повышения эффективности выявления инсайдерских угроз в корпоративных сетях.

Список литературы:

1. Greitzer F.L., Frincke D.A. Combining traditional cyber security audit data with psychosocial data: Towards predictive modeling for insider threat mitigation // Insider Threats in Cyber Security — 2010. — P. 85–113.
2. Chandola V., Banerjee A., Kumar V. Anomaly detection: A survey // ACM Computing Surveys — 2009. — Vol. 41. №3. — P. 1–58.
3. Eberle W., Holder L. Discovering structural anomalies in graph-based data // Proc. IEEE Int. Conf. Data Mining — 2007. — P. 393–398.
4. Bishop C. M. Pattern Recognition and Machine Learning — New York: Springer, 2006. — P. 537–590.
5. Aggarwal C. C. Outlier Analysis — New York: Springer, 2017. — P. 45–120.
6. Akoglu L., Tong H., Koutra D. Graph-based anomaly detection and description: A survey // Data Mining and Knowledge Discovery — 2015. — Vol. 29. №3. — P. 626–688.
7. Behl A., Behl K., Behl D. Cybersecurity and Cyberwar: What Everyone Needs to Know — Oxford: Oxford University Press, 2017. — P. 102–145.

Мингазов Т.Р., Калинин М.О.

Санкт-Петербургский политехнический университет Петра Великого, Санкт-Петербург

КОМПЛЕКСНАЯ ОЦЕНКА ЭФФЕКТИВНОСТИ АДАПТИВНОЙ ЗАЩИТЫ СИСТЕМ УДАЛЕННОГО ДОСТУПА К ХОСТИНГ-СЕРВИСАМ

Задача адаптивного управления параметрами защиты систем удаленного доступа к корпоративным хостинг-сервисам малых и средних предприятий (МСП) формализуется как стохастическая оптимизация с ограничениями [1]. В качестве интегрального критерия выступает комбинация $J(u)=E[\lambda_1 R+\lambda_2 L+\lambda_3 C]$, где R – функция риска, L – потери доступности, C – затраты, а управление $u = (\tau, \alpha, \gamma)$ объединяет период MTD-перестройки τ, параметры строгости политик ZTA/SDP α и режим защиты клиентского приложения γ. Представленное исследование направлено на разработку спецификаций функций R, L, C применительно к ограниченным ресурсам МСП и анализу возникающих между параметрами защиты перекрестных связей.

В качестве модели функции риска принимается параметризация, согласующаяся по знаку с гипотезой MTD-энтропии [2]. Рост непредсказуемости конфигурации для нарушителя снижает вероятность успешной атаки. Обозначив $H(\tau)$ энтропию доступного защитнику конфигурационного пространства за интервал τ, $p_{zt}(\alpha)$ – вероятность обхода политики ZTA/SDP (убывает с ростом α [3]) и $p_{cl}(\gamma)$ – вероятность компрометации клиентского приложения (убывает с ростом γ [4]), опишем риск как $R(u, w) = w_a \exp(-\beta \cdot H(\tau)) + (1-w_a) p_{zt}(\alpha)\, p_{cl}(\gamma)$, где $w_a \in [0,1]$ – нормированная интенсивность сетевой атаки (компонента вектора возмущений), $\beta>0$ – эмпирический коэффициент эффективности MTD

для конкретной инфраструктуры. Первое слагаемое отвечает за устойчивость к сетевым атакам (DDoS, сканирование), второе – за многоэтапное проникновение через обход контроля доступа и компрометацию клиента. Экспоненциальная свертка использована как упрощение и подлежит калибровке по реальным данным. Допустимы альтернативные параметризации, например, степенная, логистическая.

Функция потерь доступности учитывает задержку $d_{zt}(\alpha)$ на верификацию каждого запроса в ZTA/SDP-шлюзе и амортизированный простой δ/τ от MTD-переключений (переустановка SPA-сессии, обновление маршрутов). При интенсивности легитимного потока μ и применении формулы Полячека–Хинчина для очереди M/G/1 [5] $L(u) = d_{zt}(\alpha) + \delta/\tau + \rho^2(\sigma_s^2+d_{eff}^2)/[2(1-\rho)]$, где $d_{eff}=d_{zt}(\alpha)+\delta/\tau$, $\rho=\mu\cdot d_{eff}$, σ_s^2 – дисперсия времени обслуживания. Слагаемое δ/τ существенно, т.к. сокращение τ повышает $H(\tau)$, но пропорционально увеличивает среднюю задержку.

Затраты складываются из нагрузки на оркестрацию MTD, операций ZTA/SDP и механизмов защиты клиента, конкурирующих за общий вычислительный бюджет B_{max} устройства МСП: $C(u)=c_{mtd}/\tau+c_{zt}(\alpha)+c_{cl}(\gamma)$, $C(u)\le B_{max}$.

Определенные выше выражения делают явной несепарабельность задачи. Уменьшение τ одновременно снижает R, повышает L и потребляет бюджет в C. Ужесточение α снижает p_{zt}, но повышает d_{zt} во втором слагаемом и c_{zt} в третьем, а прирост c_{zt} сокращает остаток бюджета для $c_{cl}(\gamma)$, косвенно увеличивая p_{cl}. Перекрестные связи параметров защиты через общий бюджет МСП делают изолированную настройку отдельных механизмов субоптимальной, что отличает постановку от известных работ по оптимизации MTD без учета политик доступа и состояния клиента [1]. Для минимизации $J(u)$ при $L(u)\le L_{max}$, $C(u)\le B_{max}$ применяется двухконтурная декомпозиция [1] с методом Ляпунова drift-plus-penalty [6] в медленном контуре. Полученные формальные выражения служат основой для численного решения задачи и калибровки параметров на стенде Docker/Kubernetes (fwknop, hping3, mitmproxy, Frida).

Список литературы:

1. Мингазов Т.Р., Калинин М.О. Формализация задачи адаптивного управления параметрами защиты систем удаленного доступа к корпоративным хостинг-сервисам // Неделя науки ИКНК : Материалы Всероссийской научно-практической конференции. – СПб. : ПОЛИТЕХ-ПРЕСС, 2026. – С. 168-170.
2. Zhuang R., DeLoach S.A., Ou X. Towards a Theory of Moving Target Defense // Proc. 2nd ACM CCS Workshop on Moving Target Defense. – 2014. – P. 31–40.
3. Rose S., Borchert O., Mitchell S., Connelly S. Zero Trust Architecture // NIST Special Publication 800-207. – 2020.
4. Harshini T. et al. Decoding the solution for man-at-the-end attacks and reverse engineering on IoMT devices // Journal of Multidisciplinary Healthcare. – 2025. – Vol. 18. – P. 6479–6501.
5. Kleinrock L. Queueing Systems, Vol. I: Theory. – New York: Wiley, 1975. – 417 p.
6. Neely M.J. Stochastic Network Optimization with Application to Communication and Queueing Systems. – Morgan & Claypool, 2010. – 199 p.

Неаскин С.Р.
Военно-космическая академия им. А.Ф. Можайского, Санкт-Петербург

АЛГОРИТМ КЛАССИФИКАЦИИ ДЕСТРУКТИВНЫХ ВОЗДЕЙСТВИЙ В АВТОМАТИЗИРОВАННЫХ ИНФОРМАЦИОННЫХ СИСТЕМАХ НА ОСНОВЕ ОПЕРАТИВНОГО АНАЛИЗА ЖУРНАЛОВ СОБЫТИЙ

Устойчивое функционирование автоматизированных информационных систем требует своевременного обнаружения и классификации деструктивных воздействий как преднамеренного, так и непреднамеренного характера [1]. Существующие средства мониторинга и обнаружения вторжений преимущественно ориентированы на выявление целенаправленных атак, тогда как сбои оборудования, ошибки персонала и иные непреднамеренные воздействия, способные привести к сопоставимым последствиям, учитываются недостаточно [2-4]. В качестве основного источника информации для выявления таких воздействий целесообразно использовать журналы событий компонентов автоматизированной информационной системы, прежде всего журналы подсистемы аудита, фиксирующие события на уровне системных вызовов ядра операционной системы [5]. В условиях эксплуатации автоматизированных информационных систем, в том числе на базе AstraLinuxSpecialEdition, актуальной задачей является разработка алгоритмов, обеспечивающих централизованный сбор и корреляцию журналов событий, а также оперативную и точную классификацию воздействий.

Разработанный алгоритм предназначен для оперативной классификации деструктивных воздействий в автоматизированной информационной системе (АИС) и формирования практического решения, направленного на локализацию, устранение или минимизацию последствий воздействия. В качестве исходной информации используются журналы событий, поступающие от различных компонентов инфраструктуры: серверов, рабочих станций, средств защиты информации, сетевых устройств, систем управления базами данных и других элементов АИС. На первом этапе выполняются централизованный сбор и объединение журналов событий в единый поток наблюдений. Далее записи приводятся к единому формату и обогащаются контекстной информацией, включающей сведения об источнике события, типе действия, субъекте и объекте воздействия, результате выполнения операции, а также о текущем состоянии инфраструктуры.

После предварительной обработки из потока событий выделяются информативные признаки, характеризующие состояние системы и особенности наблюдаемого воздействия. К таким признакам могут относится события аутентификации и повышения привилегий, обращения к критическим файлам, запуск процессов, ошибки сервисов и приложений, изменения конфигурации, отклонения в сетевой активности, признаки аппаратных отказов, нарушения электропитания, а также иные аномалии, выявляемые на системном и инфраструктурном уровнях. Использование совокупности признаков различной природы позволяет сформировать более полное представление о происходящем в системе и учитывать, как технические, так и эксплуатационные причины инцидентов.

На основе выделенных признаков алгоритм осуществляет разделение воздействий на два основных класса: преднамеренные и непреднамеренные. После определения общего характера воздействия выполняется его последующая детализация. К преднамеренным воздействиям могут быть отнесены попытки несанкционированного доступа, повышение привилегий, изменение конфигурации, внедрение вредоносного кода, удаление или модификация данных. К непреднамеренным воздействиям относятся аппаратные и программные сбои, ошибки пользователей и администраторов, нарушения регламента эксплуатации, сбои электропитания и иные отказы инфраструктурных компонентов. Такой подход позволяет не только фиксировать факт нарушения, но и определять его природу, что имеет принципиальное значение для выбора корректного способа реагирования.

В отличие от стандартных средств обнаружения воздействий, ориентированных главным образом на фиксацию известных инцидентов и аномалий по правилам корреляции, разработанный алгоритм обеспечивает статистически обоснованную оперативную классификацию как воздействий по мере поступления событий путем последовательного анализа методом Вальда. Это особенно важно в сценариях деградации сетевого взаимодействия, проявляющихся, например, в снижении скорости передачи файла, где стандартные средства часто фиксируют лишь общий факт нарушения, но не определяют его природу.

Экспериментальная проверка разработанного алгоритма проводилась на стенде, моделирующем типовую АИС под управлением AstraLinuxSpecialEdition. Установлено, что деструктивные воздействия приводят к существенному увеличению времени передачи информации и снижению средней фактической скорости передачи данных.

В отличие от традиционных средств мониторинга, ориентированных преимущественно на фиксацию факта нарушения, предложенный подход позволяет различать преднамеренные и непреднамеренные инциденты и формировать решение, пригодное для практического реагирования.

Экспериментальные результаты подтвердили работоспособность алгоритма и его преимущество по скорости обнаружения в условиях деградации сетевого взаимодействия АИС. Практическая значимость работы состоит в возможности применения предложенного подхода для повышения устойчивости функционирования АИС за счёт ускоренного анализа журналов событий и более своевременного реагирования на инциденты [6-7].

Список литературы:

1. Малюк, С. В. Пазизин, Н. С. Погожин Введение в защиту информации в автоматизированных системах: учеб. пособие – М. : Горячая линия-Телеком, 2001. – 148 с.
2. К. В. Стародубов, Х. А. Х. Шамсулдин, М. А. Д. Аль-Амиди, А. А. Л. Алмали Анализ решений в области средств защиты информации в автоматизированной системе / // Проектное управление в строительстве. – 2023. – № 2(29). – С. 81-85.
3. Баринов, А. Е. Классификация каналов деструктивного информационного воздействия на программируемые логические контроллеры сетей АСУ ТП / А. Е. Баринов, А. М. Богер // Цифровая Индустрия: Состояние и Перспективы Развития 2023 (ЦИСП'2023) : Сборник научных статей, Челябинск, 21–23 ноября 2023 года. – Челябинск: Издательский центр Южно-Уральского государственного университета, 2024. – С. 88-94.
4. Попов, *А. М.* Устойчивость распределенной автоматизированной системы управления с учетом модели прогнозирования последствий непреднамеренных деструктивных воздействий / А. М. Попов, В. И. Филатов // Проблемы машиностроения и надежности машин. – 2024. – № 4. – С. 84-89. – DOI 10.31857/S0235711924040111.
5. Справочный центр AstraLinux. – URL:https://wiki.astralinux.ru/pages/viewpage.action ... (дата обращения: 11.02.2026).
6. Свид. о гос. регистрации программы для ЭВМ № 2026614155. Программа классификации деструктивных воздействий на автоматизированную систему (СКДВ-АС) / С. Р. Неаскин. – Заявл. 06.02.2026, опубл. 12.02.2026.
7. Свид. о гос. регистрации программы для ЭВМ № 2026613244. Программа многокритериальной оптимизации параметров безопасности с учётом вероятностных рисков и ресурсных ограничений / С. Р. Неаскин. – Заявл. 01.02.2026, опубл. 05.02.2026.

Пагуба Г.Ю., Павленко Е. Ю., Зегжда Д. П.
Санкт-Петербургский политехнический университет Петра Великого, Санкт-Петербург

ПОИСК НЕДЕКЛАРИРОВАННЫХ ВОЗМОЖНОСТЕЙ В ФОРМАЛЬНОМ ОПИСАНИИ СЛОЖНО-ФУНКЦИОНАЛЬНЫХ БЛОКОВ

Одной из наиболее актуальных угроз конфиденциальности, целостности и доступности информации, обрабатываемой встраиваемыми системами, является злонамеренное внедрение недекларированного функционала в аппаратное обеспечение, являющееся частью данных встраиваемых систем. В качестве наиболее актуальных инцидентов данного рода можно выделить обнаружение уязвимости удаленного выполнения программного кода в подсистеме Intel Management Engine [1] и использование недокументированных MMIO регистров для обхода подсистемы защиты памяти процессоров Apple в ходе выполнения вредоносной компании «Операция «Триангуляция» [2].

Документ «Сводная стратегия развития обрабатывающей промышленности Российской Федерации до 2035 года» [3] подразумевает отказ от зарубежной интеллектуальной собственности, в том числе от зарубежных сложно-функциональных (СФ) блоков. Однако, синтез отечественных СФ блоков является трудоёмкой задачей. В связи с этим, актуальной является задача поиска недекларированных возможностей в СФ блоках.

По данным источника [4] выделяют следующие методы формальной верификации аппаратного обеспечения, в том числе сложно-функциональных блоков:

1. Потактовая симуляция.
2. Функциональная верификация.
3. Тестирование с помощью ПЛИС.

Основной проблемой потактовой симуляции и тестирования с помощью ПЛИС является отсутствие гарантий полноты покрытия пространства возможных состояний тестируемого аппаратного обеспечения. Функциональная верификация же в свою очередь требует создания специализированных средств тестирования, заточенных под конкретные образцы аппаратного обеспечения.

В данной работе предлагается метод поиска недекларированных возможностей в формальном описании сложно-функциональных блоков на основе символьной интерпретации.

Для обеспечения поиска недекларированных возможностей в формальном описании СФ блоков, данное формальное описание представляется в виде кортежа $IP = < S_{input}, S_{inter}, S_{output}, C >$, где:

$S_{input} = \{s_{input_1}, s_{input_2}, \dots, s_{input_n}\}$ – множество, каждый элемент которого описывает входной сигнал СФ блока;

$S_{inter} = \{s_{inter_1}, s_{inter_2}, \dots, s_{inter_m}\}$ – множество, каждый элемент которого описывает внутренний сигнал СФ блока;

$S_{output} = \{s_{output_1}, s_{output_2}, \dots, s_{output_k}\}$ – множество, каждый элемент которого описывает выходной сигнал СФ блока;

$C = \{c_1, c_2, \dots, c_r\}$ – множество полиномов Жегалкина, описывающих булевые функции зависимостей между входными, выходными и внутренними сигналами СФ блока.

Элементы множеств S_{input}, S_{inter} и S_{output} имеют вид $s_i = \{name, type, size\}$, где:

$name$ – имя сигнала в формальном описании СФ блока;

$type$ – тип сигнала ($wire, logic$ и др.);

$size$ – максимальный размер битового вектора, способного хранить состояние описываемого сигнала.

Предлагаемый метод имеет следующие основные этапы выполнения:

1. Анализ формального описания СФ блока. На данном этапе происходит анализ формального описания СФ блока и выделение основных сущностей.
2. Представление СФ блока в терминах предлагаемой модели (составление кортежа *IP*).
3. Составление полиномов Жегалкина, описывающих зависимости между сигналами СФ блока.
4. Получение пространства возможных состояний СФ блока посредством разрешения составленных ранее зависимостей между сигналами СФ блока.
5. Определение различий между полученным пространством состояний и декларированным пространством состояний.

Предлагаемый метод позволяет осуществлять автоматизированный поиск недекларированных возможностей в формальном описание СФ блоков и уменьшает трудоемкость выполнения задач данного рода.

Список литературы:

1. BlackHat. How To Hack A Turned-Off Computer, Or Running Unsigned Code In Intel ME [Электронный ресурс]. URL: https://blackhat.com/docs/eu-17/materials/eu-17-Goryachy-How-To-Hack-A-Turned-Off-Computer-Or-Running-Unsigned-Code-In-Intel-Management-Engine-wp.pdf. – (дата обращения: 10.05.2026).
2. SecureList. Операция «Триангуляция»: последняя аппаратная загадка [Электронный ресурс]. URL: https://securelist.ru/operation-triangulation-the-last-hardware-mystery/108683/. – (дата обращения: 10.05.2026).
3. Правительство России. Сводная стратегия развития обрабатывающей промышленности Российской Федерации до 2035 года [Электронный ресурс]. URL: http://static.government.ru/media/files/AIAVFpbzBo7cvkwaMoNtWjJLt6WA8Cmu.pdf. – (дата обращения: 10.05.2026).
4. OWASP Mobile Top 10 [Электронный ресурс]. URL: https://owasp.org/www-project-mobile-top-10/. – (дата обращения: 12.05.2026).

Пеньков Д.В., Крюков Р.О.

Военно-космическая академия имени А.Ф. Можайского, Санкт-Петербург

ПОДХОД К МОНИТОРИНГУ БЕЗОПАСНОСТИ РАСПРЕДЕЛЕННЫХ ПРИЛОЖЕНИЙ НА ОСНОВЕ ИНТЕГРАЦИИ ТЕХНОЛОГИЙ EBPF И OPENTELEMETRY

Введение. Современные высоконагруженные распределённые информационные системы, развёртываемые в средах контейнерной оркестрации (Kubernetes), порождают принципиальный структурный разрыв в средствах мониторинга безопасности. Системы управления информацией и событиями безопасности (SIEM, Security Information and Event Management) ориентированы на обработку системных журналов уровня операционной системы, тогда как инструменты управления производительностью приложений (APM, Application Performance Management) фиксируют метрики и трассы прикладного уровня. Изолированное функционирование указанных классов решений не обеспечивает восстановления целостной картины инцидента: в частности, не позволяет соотнести запуск вредоносного процесса в контейнере (системное событие) с инициировавшим его HTTP-запросом (прикладной контекст) [1, 2].

Основная часть. Технология eBPF (extendedBerkeleyPacketFilter) обеспечивает безагентный мониторинг ядра Linux и перехват системных вызовов (execve, connect, open) без модификации исходного кода целевых приложений и без демаскирующих признаков

для потенциального нарушителя [3]. Вместе с тем события, регистрируемые eBPF-агентами (Falco, Tetragon), не содержат прикладного контекста и не позволяют идентифицировать микросервис либо пользователя, инициировавшего вызов. Стандарт OpenTelemetry (OTel), являющийся де-факто открытым стандартом сбора трасс, метрик и журналов в распределённых системах, описывает прохождение запроса по микросервисам, однако не регистрирует событий уровня операционной системы [4].

Научная новизна предлагаемого подхода заключается в формировании единой аналитической плоскости -SecurityObservability- посредством интеграции технологий eBPF и OpenTelemetry. В отличие от существующих коммерческих решений (Sysdig, AquaSecurity), в которых eBPF применяется изолированно либо посредством закрытых проприетарных агентов, разработанная архитектура впервые реализует обогащение системных событий прикладным контекстом через открытый стандарт OpenTelemetry. Указанный подход обеспечивает не только фиксацию системного вызова execve ("/bin/sh"), но и его атрибуцию конкретной транзакции (trace_id), инициированной, в частности, неавторизованным клиентом через уязвимый API-эндпоинт. С целью практической реализации спроектирована архитектура, объединяющая поток событий eBPF от Cilium Tetragon с конвейером OpenTelemetryCollector (рисунок 1).

Базовым механизмом корреляции выступают общие идентификаторы среды выполнения (namespace, pod_name, PID); в качестве перспективного направления рассматривается внедрение идентификатора trace_id непосредственно в адресное пространство процессов. Разработанная система развёрнута в Kubernetes-кластере и включает: микросервисное приложение (Gateway, Auth, Order), инструментированное при помощи OTel SDK; уровень сбора системных событий (Tetragon/eBPF); центральный процессинговый узел – Open Telemetry Collector, обеспечивающий обогащение и маршрутизацию телеметрии; подсистемы хранения и визуализации -Jaeger (трассировка), Prometheus (метрики), Loki (журналы) и Grafana (единая панель оператора).

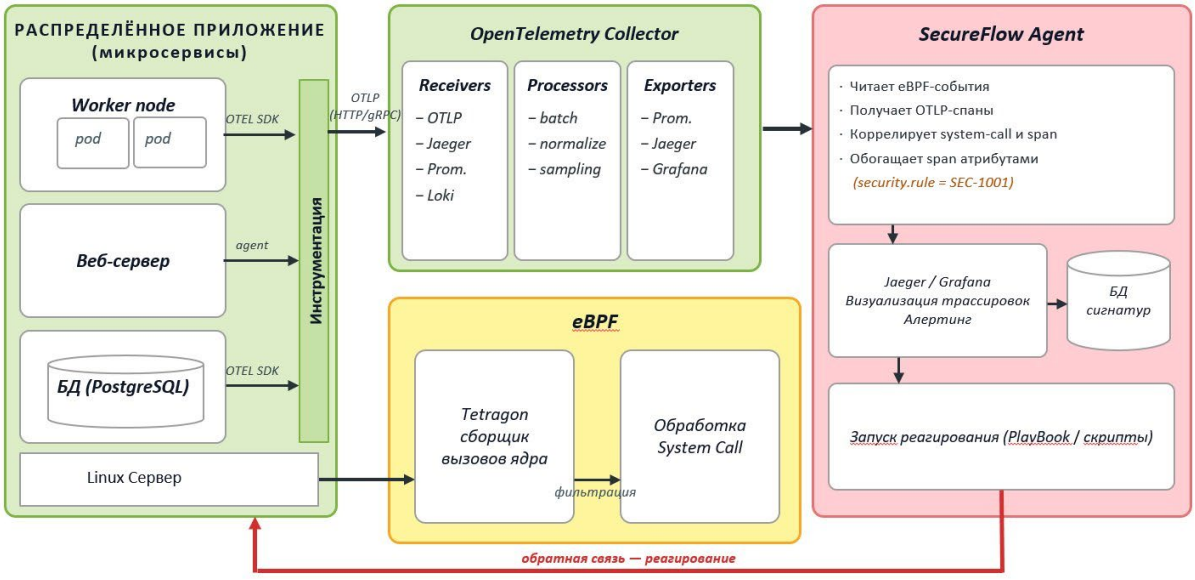


Рисунок 1 - Архитектура системы мониторинга безопасности распределённых приложений

Выводы. Предложенная архитектура трансформирует наблюдаемость распределённых приложений из пассивного диагностического инструмента в активный механизм противодействия компьютерным атакам в облачных средах. Решение является языково- и платформонезависимым, обладает свойством горизонтальной масштабируемости и совместимо с устоявшейся экосистемой DevSecOps (Grafana, Jaeger, OpenTelemetry), что обеспечивает поэтапное внедрение без замены существующей

инфраструктуры. Архитектура допускает эксплуатацию в кластерах, насчитывающих тысячи узлов, без значимой деградации показателей производительности. Полученные результаты согласуются с требованиями к мониторингу безопасности, предъявляемыми к средствам контейнеризации в отечественных информационных системах [5].

Список использованных источников

1. Cloud Native Computing Foundation. eBPF for Observability and Security in Cloud-Native Environments[Электронныйресурс]2023.URL:https://www.cncf.io/blog/2023/04/12/ebpf-for-observability-and-security-in-cloud-native-environments/ (дата обращения: 02.05.2024).
2. Isovalent. eBPF&OpenTelemetry: Bridging Kernel-Level Observability with Standardized Telemetry [Электронныйресурс], 2023. URL: https://isovalent.com/blog/post/2023-06-07-ebpf-opentelemetry/ (датаобращения: 02.05.2024).
3. Linux Foundation. eBPF-based Runtime Security and OTLP Export for Kubernetes Workloads [Электронныйресурс], 2022.URL: https://www.linuxfoundation.org/wp-content/uploads/2022/11/ebpf-otlp-security-k8s.pdf (датаобращения: 02.05.2024).
4. OpenTelemetry Community. OpenTelemetry Protocol (OTLP) Specification [Электронныйресурс], 2024. URL: https://opentelemetry.io/docs/specs/otlp/ (датаобращения: 02.05.2024).
5. Требования по безопасности информации к средствам контейнеризации, утв. приказом ФСТЭК России от 4 июля 2022 г. № 118 [Электронный ресурс]. URL: https://fstec.ru/dokumenty/vse-dokumenty/spetsialnye-normativnye-dokumenty/trebovaniya-po-bezopasnosti-informatsii-utverzhdeny-prikazom-fstek-rossii-ot-4-iyulya-2022-g-n-118 (дата обращения: 02.05.2024).

Полтавцева М.А.

Санкт-Петербургский политехнический университет Петра Великого, Санкт-Петербург

ВОПРОСЫ БЕЗОПАСНОСТИ МУЛЬТИМОДЕЛЬНЫХ ДАННЫХ

Развитие современных систем управления данными привело к появлению новых типов инструментов и экосистем обработки информации, формирование которых находится сегодня на стадии активного развития и применения. Можно выделить несколько тенденций, значительно влияющих на ландшафт инструментов в данной области:

1. Специализация.
2. Мультимодельность.

Причем, до определенной степени это две связанные тенденции. На протяжении всей эволюции СУБД разработчики искали с одной стороны более производительные решения, а с другой – более дешевые. Причем на совокупную стоимость влияли как стоимость оборудования, инфраструктуры, так и стоимость найма персонала, на которой сказывались требования к квалификации, число специалистов. В процессе развития реляционных СУБД основными трендами для них стали:

- универсальность (как основное преимущество и способ удешевления использования);
- расширение функциональности для решения все большего числа и типов задач;
- повышение качества автоматизации процессов выполнения и оптимизации запросов.

Автоматическая оптимизация запросов, позволившая реляционным СУБД занять доминирующее положение, долго не позволяло другим решениям составить им значимое преимущество. Однако, появление большого объема данных, превосходившего

возможности одного узла обработки, с одной стороны породило явление «машин баз данных», а с другой – в силу увеличения стоимости этих «машин», желание использовать для обработки данных простые и дешевые вычислительные узлы, пусть и заплатив некоторыми преимуществами. Так в начале 2000-х вместе с ростом числа дешевых вычислительных узлов (и их возможностей) появились не реляционные модели данных и No-SQL решения. Простота шаржирования, фрагментации в рамках вычислительного кластера позволили таким продуктам решать большое число задач, сложных для «монолитных» массивно-параллельных реляционных СУБД. Ценой таких решений стала универсальность. NoSQL решения хорошо справлялись с относительно узкими классами задач, для которых были предназначены. Так сложились два тренда:

1. Тренд на дешевые решения специализированных задач, который привел к появлению целых гетерогенных экосистем, использованию наборов специализированных решений в рамках единых комплексов.
2. Тренд на расширение универсальных СУБД и адаптацию к решению новых задач, включая шардирование и обработку больших данных.

Несмотря на то, что технологическая сложность решаемых задач заставила второй тренд проявиться немного позднее, сегодня мы можем ясно констатировать оба: гетерогенные системы управления данными как архитектуры инженерии данных крупных организаций и мультимодельные «коробочные» системы, являющиеся, как правило, в той или иной степени развитием традиционных реляционных СУБД. Встраивание в эти тенденции векторных СУБД для систем искусственного интеллекта уверенно подтверждает выделенные тренды [1,2]. В результате сегодня можно выделить три различных типа инструментов в области СУБД: мультимодельные СУБД, гетерогенные системы управления большими данными и полихранилища.

В любом случае, специалиста по информационной безопасности этот эволюционный процесс привел к одной и той же проблеме: обеспечение безопасности мультимодельных данных. В независимости от того, идет ли речь о «коробочном» решении или гетерогенной экосистеме, возникает общий список проблем, в отношении которых теоретические подходы становятся схожи, в отличии от имплементации в конкретные инструменты. Новыми задачами для всех типов мультимодельных систем стали:

1. Обеспечение контроля доступа в условиях различной грануляции данных.
2. Оценка защищенности с учетом не только распределенности, но и трансформации данных (включая особенности контроля дсотупа).

Очевидным образом, эти задачи сказались и на решении задачи аудита, и других смежных задачах – хотя они, в большей степени отличаются имплементацией в различные инфраструктуры, чем в теоретическом аспекте для различных типов мультимодельных систем. Ключевой проблемой обеспечения безопасности мультимодельных систем с учетом их особенностей является согласованное представление данных [3,4]. Если на базовом уровне для разработки унифицированных политик безопасности достаточно использования протсых моделей согласования гетерогенных мультимодельных данных [4], то в случае семантического анализа и необходимости целевого задания политики безопасности со стороны бизнес логики требуется более сложных подход к описанию данных, основанный на моделях концептуального уровня учитывающих семантику. Такие работы сегодня актуальны не только в безопасности, но и в задачах проектирования и управления хранилищами данных [5]. Сравнение и анализ концептуальных моделей, предлагаемых для проектирования и управления мультимодельными схемами [6] позволяет выбрать наиболее подходящие для решения задач безопасности и интегрировать их с инструментами безопасности отдельных мультимодельных платформ (мультимодельных СУБД, полихранилищ или гетерогенных систем управления большими данными).

Реализация системы контроля доступа и управления политиками безопасности, интегрированная в платформу управления мультимодельными данными, позволяет перенести управление доступом на уровень бизнес-логики, однако требует дальнейшей

интеграции с более низкоуровневыми инструментами и моделями для решения всего спектра задач безопасности: мониторинга процессов обработки данных, аудита, оценки защищенности. Актуальными остаются вопросы также согласования грануляций данных и адаптации логики поиска оптимальной политики безопасности в мультимодельной системе.

Библиографический список:

1. Lu Y., Tang H. Multimodal data storage and retrieval for embodied ai: A survey //arXiv preprint arXiv:2508.13901. – 2025.
2. Wienholt N. Databases and AI //GitHub Copilot and AI Coding Tools in Practice: Accelerate AI Adoption from Individual Developers to Enterprise. – Berkeley, CA : Apress, 2025. – С. 189-200.
3. Qu Y. et al. Database Schema Matching Model Based on Multimodal Fusion Analysis //2025 5th Asia Conference on Information Engineering (ACIE). – IEEE, 2025. – С. 32-36.
4. Poltavtseva, M. A. Conceptual Data Modeling Using Aggregates to Ensure Large-Scale Distributed Data Management Systems Security / M. A. Poltavtseva, M. O. Kalinin // Studies in Computational Intelligence. – 2020. – Vol. 868. – P. 41-47. – DOI 10.1007/978-3-030-32258-8_5. – EDN TNJKFN.
5. Geisler S. et al. Conceptual modeling of user perspectives–From data warehouses to alliance-driven data ecosystems //Data & Knowledge Engineering. – 2025. – С. 102502.
6. Paraskevas K. Data integration and storage strategies in heterogeneous analytical systems: architectures, methods, and interoperability challenges //Information. – 2025. – Т. 16. – №. 11. – С. 932.

Романов И.А., Карасёв В.Т.

Военно-космическая академия имени А.Ф. Можайского, Санкт-Петербург

СИСТЕМА ИНТЕЛЛЕКТУАЛЬНОЙ БЕЗОПАСНОСТИ УМНОГО ДОМА, ИСПОЛЬЗУЮЩАЯ АРХИТЕКТУРУ НУЛЕВОГО ДОВЕРИЯ И ПРИНЦИП ОБЪЯСНИМОСТИ

В условиях роста числа IoT-устройств в российских домохозяйствах и санкционных ограничений на облачные сервисы вопросы локальной безопасности «Умного дома» выходят на первый план. Большинство решений используют зарубежные облака, создавая риски утечек данных. Переход на edge-модели, не требующие постоянного подключения к мощным серверам, является критической задачей [1].

Умные дома становятся сложнее, а генерируемые данные – чувствительнее. Традиционные модели безопасности, основанные на доверии ко всему внутри сети, неэффективны. Архитектура нулевого доверия (ZTA) с принципом «Никогда не доверяй, всегда проверяй» требует постоянной аутентификации. Однако ZTA создает вычислительную нагрузку и задержки на граничных устройствах.

Предлагается использовать фреймворк POIZE – комплексная структура безопасности, фокусирующаяся на конфиденциальности, оптимизации потоков данных, использовании Tiny Machine Learning (TinyML) для обнаружения аномалий, применении ZTA и обеспечении объяснимости для пользователей.

Структура POIZE [2]

Предлагаемая структура состоит из ключевых слоев, обеспечивающих сквозную безопасность при ограниченных ресурсах.

1. Слой нормализации и дедупликации данных (оптимизация) направлен на снижение затрат из-за постоянной проверки ZTA. Слой удаляет избыточные данные: сравнивает новые показания сенсоров с буфером (при малой разнице данные отбрасываются) и использует хеш-суммы для отбраковки одинаковых видео-кадров.

2. Слой аутентификации (Zero Trust) использует комбинацию MFA, PKI для идентификации устройств, TLS для шифрования и OAuth 2.0 для сторонних сервисов. Проверка происходит непрерывно.

3. Слой обнаружения аномалий. Используется модель Random Forest на базе TinyML. Выбор данного метода обусловлен низкими требованиями к ресурсам по сравнению с глубокими нейронными сетями. Это критически важно для российских реалий, где пользователи часто применяют маломощные одноплатные ПК (Raspberry Pi, Repka Pi). Модель способна выполняться на устройствах типа NVIDIA Jetson Nano.

4. Слой объяснимого искусственного интеллекта (XAI) и приватности. Перед отправкой данных сторонним сервисам применяются фильтры для удаления персональной информации (PII). Слой XAI использует метод LIME для реконструкции логики решений искусственного интеллекта, генерируя понятные пользователю описания угроз.

Ключевые показатели эффективности при использовании POIZE [2]

1. Дедупликация – экономия трафика в дневное время составляет около 20%.

2. Точность – модель показала 4% ложноположительных и 5% ложноотрицательных срабатываний. AUC-ROC кривая – 0,958 (обучение) и 0,95 (тест). AUC-ROC кривые для обучающей, тестовой и контрольной выборок представлены на рисунке 1.

3. Производительность – POIZE создает накладные расходы с коэффициентом 1,8x, но это компенсируется дедупликацией. Задержка в сотни миллисекунд незаметна для пользователя.

4. Масштабируемость – рост времени отклика линеен, модель пригодна для крупных сред.

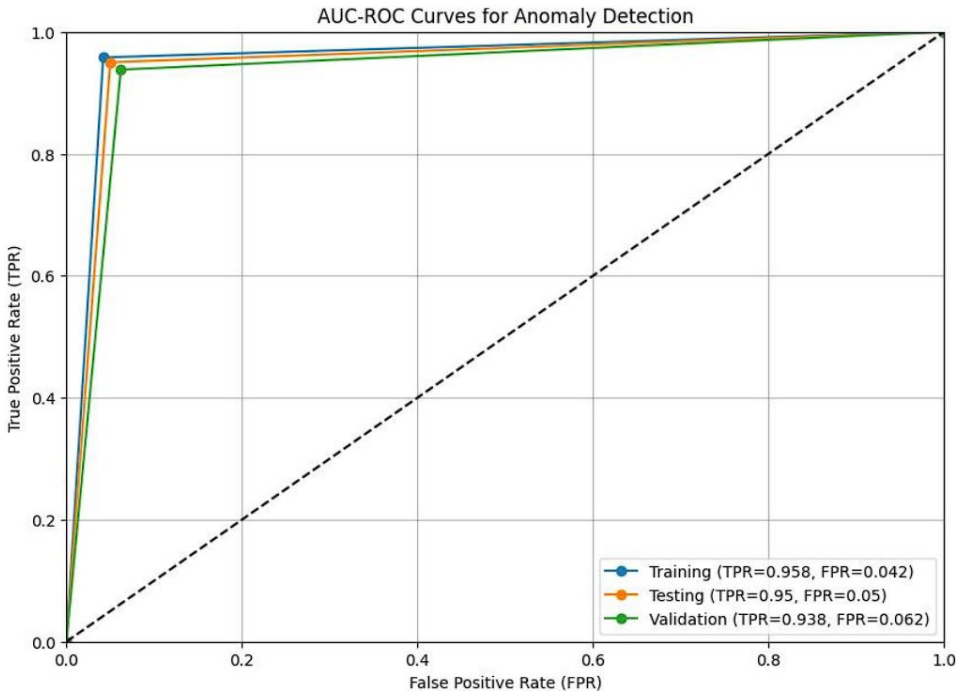


Рисунок 1. - Кривые AUC-ROC для обнаружения аномалий

Работа демонстрирует подход, потенциально адаптируемый под российское ПО и аппаратное обеспечение.

Пути решения на базе POIZE:

1. Локальный edge-вычислитель – устраняет зависимость от зарубежных облаков.

2. Низкие требования к железу – работа на доступных в РФ одноплатных ПК (Raspberry Pi, Repka Pi).

3. Объяснимость [3] – соответствие требованиям 152-ФЗ о прозрачности алгоритмов.

4. Дедупликация трафика – экономия до 50% данных, критично для регионов с нестабильным интернетом.

Направления дальнейших исследований

Требуется валидация фреймворк POIZE на реальных тестовых стендах, а не только на синтетических данных AWS. Нормализация данных от устройств разных вендоров может потребовать усилий по интеграции. Требует исследования вопрос запуска фреймворка на сверхмаломощных микроконтроллерах (ESP32), популярных в российском DIY-сегменте.

Таким образом, фреймворк POIZE обеспечивает принципы ZTA на всех уровнях, сохраняет конфиденциальность и оптимизирует потоки данных. Он способен удалять до 20% избыточного трафика, работая на маломощных устройствах. Предлагаемый подход имеет высокую практическую значимость для развития суверенных технологий «Умного дома»
в Российской Федерации.

Список литературы:

1. Технологии искусственного интеллекта: учебник / авторский коллектив: Д.Н.Бирюков, С.В.Войцеховский, В.В.Ефимов [и др.]; под общей редакцией В.В.Ефимова, А.Д.Хомоненко. – СПб: ВКА имени А.Ф. Можайского, 2024. – 1 CD-ROM. – Текст: электронный.
2. Gupta A., Gupta S., Sharma S. et al. A privacy preserving optimized intelligent security framework for smart homes using zero trust architecture and explainability. Sci Rep (2026). DOI: 10.1038/s41598-026-44587-1.
3. Moss J., Gordon J., Duclos W. et al. Explainable AI in IoT: A Survey of Challenges, Advancements, and Pathways to Trustworthy Automation. Electronics. 2025;14(23):4622. DOI:10.3390/electronics14234622.

Рохман Н.А., Александрова Е.Б.

Санкт-Петербургский политехнический университет Петра Великого, Санкт-Петербург

ОБРАТИМОЕ КОДИРОВАНИЕ ТЕКСТОВЫХ СООБЩЕНИЙ В MIDI-ДАННЫЕ НА ОСНОВЕ ПОТОКОВЫХ НЕЙРОННЫХ СЕТЕЙ

В настоящее время генеративные методы бесконтейнерной стеганографии рассматриваются как перспективное направление защиты информации, поскольку позволяют формировать медиаданные непосредственно из скрываемого сообщения без использования модифицируемого контейнера [1, 2]. Большинство исследований ориентированы преимущественно на изображения и видеоданные, тогда как применение подобных подходов к музыкальным данным изучено значительно меньше.

В работе рассматривается задача обратимого преобразования текстовых сообщений в MIDI-последовательности с использованием потоковых нейронных сетей при минимизации потерь информации на этапе преобразования латентного представления в MIDI-данные. Разработанный подход основан на использовании модели RealNVP [3] совместно с моделью автоэнкодера (Encoder-Decoder), обеспечивающими двустороннее преобразование данных в цикле (рис. 1).

Рис. 1. Схема цикла преобразования

Для подготовки музыкальных данных использовались MIDI-файлы датасета MAESTRO v3.0.0 [4]. Каждое MIDI-событие описывалось параметрами высоты ноты, громкости и временного сдвига, которые после нормализации объединялись в вектор фиксированной размерности.

Архитектура RealNVP была реализована на основе маскированных аффинных преобразований. Для каждого слоя использовались отдельные нейронные сети, включающие линейные слои, функции активации LeakyReLU и LayerNorm. Размерность входного пространства составляла 384, что соответствует 128 MIDI-событиям по три параметра на каждое событие.

На начальном этапе исследования отдельно проверялась корректность модуля кодирования и декодирования текста без использования музыкального канала. Первоначально точность восстановления текста в цикле $text \rightarrow latent \rightarrow text$ составляла 85-90%. После изменения процедуры обучения, использования датасета строк фиксированной длины и введения строгой валидации удалось добиться полного восстановления текстовых сообщений, что позволило исключить модуль преобразования текста как источник основных потерь.

Далее была выполнена проверка обратимости модели RealNVP. Для оценки использовалась метрика максимального отклонения между исходным вектором и результатом двойного преобразования $x \rightarrow g(f(x))$. Полученные значения порядка 10^{-5} показали, что модель сохраняет обратимость в пределах вычислительной точности и практически не вносит искажений в латентное пространство.

Однако при подключении музыкального канала и использовании стандартного MIDI-представления с одной нотой на временной шаг точность восстановления текста после полного цикла преобразования снижалась до 70%. Анализ ошибок показал, что основным источником потерь является параметр delta time. В процессе преобразования часть значений выходила за допустимые диапазоны MIDI и подвергалась клипированию при записи файла, что приводило к необратимому искажению латентного представления.

Для снижения потерь предлагается схема кодирования с двумя одновременно воспроизводимыми нотами на один временной шаг. Первая нота используется для хранения основного музыкального представления, включая параметры note, velocity и delta time. Вторая нота служит для хранения остаточной информации (residual-компонент) и имеет нулевой временной сдвиг, благодаря чему воспринимается как часть аккорда и не нарушает музыкальную структуру MIDI-последовательности.

В рамках residual-кодирования дополнительные компоненты высоты ноты сохраняются во второй ноте, а остаточная информация параметра delta time кодируется через параметр velocity второй ноты. Такой подход позволяет существенно уменьшить необратимые искажения, возникающие при ограничении диапазонов MIDI-значений.

Дополнительно проведена эмпирическая настройка масштабных коэффициентов латентного пространства и residual-компонент. Эксперименты показали, что оптимальная конфигурация достигается при уменьшенном масштабе латентного представления и асимметричном масштабировании residual-параметров. В данной конфигурации точность восстановления текста достигла 96,9%.

Таким образом, переход от однонотного к двухнотному представлению существенно снижает потери информации, возникающие при преобразовании $latent \rightarrow MIDI$. При этом сохраняется музыкальная интерпретируемость генерируемых последовательностей и статистическая структура исходных MIDI-данных.

Дополнительно было проведено тестирование устойчивости метода к различным преобразованиям MIDI-файлов. Установлено, что система устойчива к операциям, не затрагивающим полезную нагрузку напрямую, однако модификации параметров времени, динамики и высоты нот существенно снижают точность восстановления сообщения.

Список литературы:

1. Zhou Z. L., Cao Y., Sun X. M. Coverless information hiding based on bag-of-words model of image // Journal of Applied Sciences. – 2016. – Vol. 34, № 5. – P. 527–536.
2. Li Q., Wang X., Wang X., Ma B., Wang C., Shi Y. An encrypted coverless information hiding method based on generative models // Information Sciences. – 2021. – Vol. 553. – P. 19–30.
3. Dinh L., Sohl-Dickstein J., Bengio S. Density estimation using Real NVP [Электронный ресурс]. – Режим доступа: https://arxiv.org/abs/1605.08803 (дата обращения: 29.05.2026).
4. Hawthorne C. et al. Enabling Factorized Piano Music Modeling and Generation with the MAESTRO Dataset // International Conference on Learning Representations (ICLR). – 2019.

Сикарев И.А.[1], Абрамова А.Л. [1], Копцева Е.П. [2], Абрамов В.М.[2]

[1]Российский государственный гидрометеорологический университет, Санкт-Петербург

[2]Государственный университет морского и речного флота имени адмирала С.О. Макарова, Санкт-Петербург

КОМПЛЕКСИРОВАНИЕ МЕТОДОВ ЗАЩИТЫ ОТ КИБЕРУГРОЗ ДЛЯ СИСТЕМ ГЕОИНФОРМАЦИОНОЙ ПОДДЕРЖКИ УПРАВЛЕНИЯ БЕЗОПАСНОСТЬЮ СУДОХОДСТВА

Геоинформационная поддержка (ГИП) управления безопасностью судоходства (УБС) является быстро развивающимся направлением цифровизации деятельности водного транспорта (ДВТ) [1 - 4]. Использование в составе ГИП УБС подсистем телекоммуникаций (ПСТК) сетевого характера приводит к появлению киберугроз для ДВТ в целом. В зависимости от вида ПСТК киберугрозы могут включать атаки на навигационные системы. Одной из наиболее опасных является подмена сигнала спутниковой навигации (GNSS spoofing). GNSS (Global Navigation Satellite System) объединяет такие системы, как GPS, ГЛОНАСС и другие, и используется для точного определения координат. При GNSS spoofing злоумышленник генерирует ложный сигнал, который приводит к тому, что на экране системы управления мостиком судна отображаются ложные координаты. Аналогичная угроза существует и для системы автоматической идентификации (AIS spoofing), когда передаются поддельные данные о положении или курсе судна.

Помимо атак на навигационные каналы, серьёзную угрозу представляет Ransomware на портовую и логистическую инфраструктуру. Ransomware — это вредоносная программа, которая блокирует данные или систему и требует выкуп.

Под киберугрозами в данной работе понимаются потенциальные или реализованные действия злоумышленников, направленные на нарушение конфиденциальности, целостности и доступности информационных ресурсов транспортных систем. К таким угрозам относятся несанкционированный доступ к сетевой инфраструктуре, распространение вредоносного

программного обеспечения, атаки типа «отказ в обслуживании» (Distributed Denial of Service, DDoS), перехват и модификация передаваемых данных, а также атаки на системы управления и навигации. Особую роль в транспортных системах играют как информационные технологии (Information Technology, IT), так и операционные технологии (Operational Technology, OT), управляющие физическими процессами, что делает инфраструктуру транспорта особенно чувствительной к кибератакам.

Несмотря на значительное количество существующих решений в области кибербезопасности, проблема комплексирования методов защиты от киберугроз для ГИП УБС остается недостаточно изученной. В большинстве исследований рассматриваются отдельные технологии или конкретные типы угроз, тогда как комплексный сравнительный анализ методов защиты с количественной оценкой их эффективности и визуализацией результатов применяется относительно редко. Отсутствие единого подхода к оценке эффективности мер кибербезопасности затрудняет выбор оптимальных решений для защиты транспортной инфраструктуры. В связи с этим актуальной научной задачей является комплексирование методов защиты, а также разработка методики сравнительного анализа методов защиты от киберугроз для ГИП УБС с использованием количественных показателей эффективности и средств визуализации результатов. Применение методов визуального анализа позволяет более наглядно представить влияние различных защитных механизмов на снижение уровня риска и облегчает принятие решений при проектировании систем кибербезопасности для ГИП УБС.

Целью данной работы является комплексирование методов защиты от киберугроз на основании сравнительного анализа эффективности методов защиты от киберугроз в транспортных системах. Для достижения поставленной цели в работе рассматриваются основные типы киберугроз в транспортной инфраструктуре, анализируются современные методы защиты информационных и коммуникационных систем транспорта, а также предлагается модель количественной оценки эффективности защитных мер с использованием показателей снижения риска.

В ходе исследования использовались следующие методы и технологии:

--системный анализ области исследования,

- технологии управления информационной безопасностью систем управления.

Рассмотрим основные результаты исследования. Установлено, для обеспечения устойчивой и безопасной ДВТ требуется комплексный подход к защите от киберугроз для ГИП УБС. Указано, что на практике применяются технические, криптографические и организационные методы информационной защиты (ИЗ), каждый из которых решает определённый класс задач.

К техническим методам относятся межсетевые экраны, системы обнаружения и предотвращения вторжений (Intrusion Detection System / Intrusion Prevention System — IDS/IPS), сегментация сетевой инфраструктуры и системы мониторинга безопасности. К криптографическим методам относятся технологии защищенной передачи данных, включая виртуальные частные сети (Virtual Private Network, VPN), протоколы защищенного соединения Transport Layer Security (TLS) и современные стандарты беспроводной безопасности, такие как Wi-Fi Protected Access 3 (WPA3). Организационные методы включают разработку политик информационной безопасности, обучение персонала и внедрение систем централизованного мониторинга и реагирования на инциденты безопасности (Security Operations Center, SOC).

Установлено, что современные киберугрозы для ДВТ затрагивают как программно-аппаратные компоненты, так и персонал информационных сетей, поэтому эффективная защита от киберугроз для ГИП УБС невозможна при использовании только одного метода безопасности.

Выполнена сравнительная характеристика основных методов защиты от киберугроз для ГИП УБС с использованием критериев эффективности по различным показателям, включая стоимость, надежность, сложность реализации. Показано, что совокупность применяемых методов защиты от киберугроз для ГИП УБС способна дать высокую степень

надежности всей системы ГИП УБС, но стоимость такой комплексной защиты может оказаться достаточно высокой. Установлено, что внедрение и администрирование комплексной защиты от киберугроз для ГИП УБС может являться не тривиальной задачей.

Следует отметить, что результаты исследования обладают значительной научной новизной и могут быть использованы в различных направлениях [5 - 11], включая образовательные программы [12].

Платформа https://www.researchgate.net/profile/Valery_Abramov2/. использована в качестве инструмента коммуникации.

Список литературы:

1. Инфокоммуникационный инструментарий для управления природными рисками при мореплавании автономных судов в Арктике при изменении климата / И. А. Сикарев, В. М. Абрамов, К. С. Простакевич [и др.] // Проблемы информационной безопасности. Компьютерные системы. – 2024. – № 1(58). – С. 110-120. – DOI 10.48612/jisp/v28t-z3kr-nrn2. – EDN RUESZV.
2. Автоматизация деятельности водного транспорта с использованием факторного подхода к управлению рисками / И. А. Сикарев, А. Л. Абрамова, К. С. Простакевич [и др.] // Проблемы информационной безопасности. Компьютерные системы. – 2026. – № 1(68). – С. 168-175. – DOI 10.66424/2071-8217-2026-1-12. – EDN STKTVM.
3. Сикарев, И. А. Аспекты разработки и дальнейшие перспективы программы автоматической обработки спутниковых архивов гидрохимических данных на языке программирования Python / И. А. Сикарев, А. И. Честнов, В. М. Абрамов // Проблемы информационной безопасности. Компьютерные системы. – 2022. – № 4(52). – С. 101-109. – DOI 10.48612/jisp/7747-zanp-m6dr. – EDN NATOGA.
4. Цифровизация защиты информации в телекоммуникационных подсистемах полуавтономных надводных судов с внешним экипажем / И. А. Сикарев, В. М. Абрамов, К. С. Простакевич [и др.] // Методы и технические средства обеспечения безопасности информации. – 2024. – № 33. – С. 88-89. – EDN GMUWDH.
5. Priorities and challenges of the state policy of the Russian Federation in the Arctic science / G. Gogoberidze, V. Abramov, E. Rumyantseva [et al.] // 17th international multidisciplinary scientific geoconference SGEM 2017, Albena, Bulgaria, 29 June – 05 July 2017. Vol. 17. – Albena, Bulgaria, 2017. – P. 721-726. – DOI 10.5593/sgem2017/52/S20.092. – EDN ETLPDM.
6. Marine economic potential assessment for environmental management in the Russian Arctic and subarctic coastal regions / G. Gogoberidze, V. M. Abramov, L. N. Karlin [et al.] // 14th international multidisciplinary scientific geoconference SGEM 2014: Conference Proceedings, Albena, 17–26 June 2014. Vol. 3. – Sofia, 2014. – P. 253-260. – DOI 10.5593/SGEM2014/B53/S21.034. – EDN LDAIAX.6.
7. Web-based tools for natural risk management while large environmental projects / E. P. Istomin, V. M. Abramov, O. M. Lepeshkin [et al.] // 19th International Multidisciplinary Scientific GeoConference SGEM 2019: Conference proceedings. ENVIRONMENTAL ECONOMICS, Albena, 30 June – 06 July 2019. Vol. 19. – Sophia, 2019. – P. 953-960. – DOI 10.5593/sgem2019/5.3/S21.120. – EDN BHHPYQ.
8. Clean technologies development strategy for the national black carbon controlling system in the Russian Arctic / V. M. Abramov, G. G. Gogoberidze, L. N. Karlin [et al.] // 14th international multidisciplinary scientific geoconference SGEM 2014, Albena, Bulgaria, 17–26 June 2014. Vol. 2. – Sofia, 2014. – P. 313-320. – DOI 10.5593/SGEM2014/B42/S19.041. – EDN YONXKJ.
9. Abramov, V. M. Innovative geoinformation technologies within management of natural risks in Venezuela / V. M. Abramov, E. P. Istomin, J. A. Garcia // 18th International Multidisciplinary Scientific GeoConference SGEM 2018: Conference proceedings,

Albena, Bulgaria, 02–08 July 2018. Vol. 18. – Albena, Bulgaria, 2018. – P. 261-268. – DOI 10.5593/sgem2018/2.2/S08.033. – EDN XKNRQX.
10. Abramov, V. M. Decadal variability of particulate matter in Moscow megacity air / V. M. Abramov, N. Popov, R. I. Bachiev // 16th International Multidisciplinary Scientific GeoConference SGEM 2016: Conference Proceedings, Albena, Bulgaria, 30 June – 06 July 2016. Vol. 2. – Albena, Bulgaria. – P. 323-330. – DOI 10.5593/SGEM2016/B42/S19.042. – EDN IHXCLN.
11. К вопросу о стратегии создания национальной системы контроля черного углерода в российской Арктике / Л. Н. Карлин, В. М. Абрамов, Г. Г. Гогоберидзе [и др.] // Ученые записки Российского государственного гидрометеорологического университета. – 2014. – № 36. – С. 67-73. – EDN TQUBAN.
12. Digital learning technologies within geo-information management / I. A. Sikarev, S. I. Lukyanov, N. Popov, [et al.] // E3S Web of Conferences, Chelyabinsk, 17–19 February 2021. – Chelyabinsk, 2021. – P. 01004. – DOI 10.1051/e3sconf/202125801004. – EDN GWVYAN.

Сикарев И.А.[1], Абрамова А.Л. [1], Простакевич К.С. [1], Абрамов В.М.[2]
[1]Российский государственный гидрометеорологический университет, Санкт-Петербург
[2]Государственный университет морского и речного флота имени адмирала С.О. Макарова, Санкт-Петербург

АСПЕКТЫ ИНФОРМАЦИОННОЙ БЕЗОПАСНОСТИ ПРИ ИСПЫТАНИЯХ АВТОНОМНЫХ И ПОЛУАВТОНОМНЫХ МАЛОМЕРНЫХ СУДОВ

Глобальный сектор разработки автономных и полуавтономных маломерных судов (АПАМС) в последние годы характеризуется активными водными испытаниями прототипов [1-4]. Очевидно, что подобные испытания, должны осуществляться с учетом всех аспектов требований безопасности, включая информационную безопасность (ИБ) [12-17]. Вместе с тем, на современном этапе существует лишь фрагментарное освещение аспектов ИБ в ходе испытаний в АПАМС. В настоящее время и имеется необходимость углубленной разработки концептуальных аспектов ИБ в ходе системных испытаний для прототипов АПАМС. Системный анализ области исследований выявил, что проводимые испытания АПАМС проводятся по упрощенным программам, в ходе которых тестированию подвергаются, в основном, системы управления движением и навигации. Следует отметить, для успешного внедрения АПАМС значительную роль играет выявление в ходе испытаний различных аспектов ИБ прототипов АПАМС, в том числе информационной защищенности процесса обновления их систем управления.

Целью работы является разработка схемы обеспечения ИБ процесса обновления системы управления АПАМС по беспроводному каналу связи (БПКС), позволяющей гарантировать аутентичность, целостность и авторизацию обновляемого программного обеспечения (ОПО).

Следует отметить, что рассмотрение испытаний так называемых одноразовых АПАМС в задачи настоящих исследований не входит.

В ходе исследований применялись:

- теория эксперимента (ТЭ);
- методы разработки защищенных автоматизированных систем (ЗАС).

Перейдем к результатам исследований. Распространённым решением защиты беспроводного обновления является применение протокола безопасности транспортного уровня (TLS, Transport Layer Security) для шифрования канала передачи вместе с проверкой

хеш-суммы пакета на стороне устройства. Для АПАМС целесообразно использовать подпись пакетов обновлений на основе алгоритма цифровой подписи на эллиптических кривых (ECDSA, Elliptic Curve Digital Signature Algorithm). Верификация опирается на доступность удалённого сервера и проверки сертификатов, что делает её неработоспособной при отсутствии связи. Общим недостатком перечисленных подходов является то, что они обеспечивают защиту лишь на уровне канала передачи, оставляя незащищёнными сценарии компрометации самого сервера дистрибуции, отката к уязвимой версии программного обеспечения и длительной автономной работы аппарата вне зоны связи.

В разработанной в ходе исследований схеме загрузки ОПО управления АПАМС по БПКС верификация подлинности и целостности обновления выполняется полностью на борту. Основу составляет трёхуровневая инфраструктура открытых ключей (PKI, Public Key Infrastructure): корневой удостоверяющий центр (Root Certificate Authority) производителя АПАМС, промежуточный центр подписи прошивок и листовые сертификаты конкретных сборок программного обеспечения. Корневой удостоверяющий центр работает вне сети — его закрытый ключ хранится в аппаратном модуле безопасности (HSM, Hardware Security Module), используется для выпуска промежуточных сертификатов. Каждый пакет обновления содержит бинарный образ прошивки, манифест с метаданными, цифровую подпись алгоритмом ECDSA P-384 и полную цепочку сертификатов.

Корневые ключи верификации прошиваются в однократно программируемые ячейки памяти на этапе производства и недоступны для перезаписи в ходе эксплуатации, что обеспечивает работу протокола безопасной загрузки (Secure Boot): при каждом старте системы каждый компонент загрузчика верифицирует подпись следующего перед его исполнением. Для защиты от отката к уязвимой версии, в манифест вносится монотонный счётчик версий, хранящийся в защищённой от перезаписи области памяти; установка пакета с номером версии ниже текущего значения счётчика блокируется аппаратно.

Предлагаемая модель включает три функциональных уровня. Уровень производителя содержит защищённый аппаратный модуль безопасности для подписи прошивок и конвейер непрерывной интеграции и доставки (CI/CD, Continuous Integration/Continuous Delivery) с автоматической генерацией манифестов. Транспортный уровень реализует туннель взаимной аутентификации (mTLS, mutual TLS), при которой и сервер, и бортовой модуль предъявляют клиентские сертификаты, исключая подключение неавторизованных источников. Бортовой уровень обеспечивает изолированный менеджер обновлений, аппаратный верификатор подписи и двухслотовую схему хранения образов: новое обновление записывается в резервный слот и активируется лишь после успешной верификации при загрузке, при отказе — система автоматически возвращается к предыдущей версии.

Результаты исследования обладают значительной научной новизной и могут быть использованы в различных направлениях [5 - 11], включая образовательные программы [12].

Платформа https://www.researchgate.net/profile/Valery_Abramov2/. использована в качестве инструмента коммуникации.

Список литературы:

1. Цифровизация защиты информации в телекоммуникационных подсистемах полуавтономных надводных судов с внешним экипажем / И. А. Сикарев, В. М. Абрамов, К. С. Простакевич [и др.] // Методы и технические средства обеспечения безопасности информации. – 2024. – № 33. – С. 88-89. – EDN GMUWDH.
2. Автоматизация деятельности водного транспорта с использованием факторного подхода к управлению рисками / И. А. Сикарев, А. Л. Абрамова, К. С. Простакевич [и др.] // Проблемы информационной безопасности. Компьютерные системы. – 2026. – № 1(68). – С. 168-175. – DOI 10.66424/2071-8217-2026-1-12. – EDN STKTVM.

3. Сикарев, И. А. Аспекты разработки и дальнейшие перспективы программы автоматической обработки спутниковых архивов гидрохимических данных на языке программирования Python / И. А. Сикарев, А. И. Честнов, В. М. Абрамов // Проблемы информационной безопасности. Компьютерные системы. – 2022. – № 4(52). – С. 101-109. – DOI 10.48612/jisp/7747-zanp-m6dr. – EDN NATOGA.
4. Инфокоммуникационный инструментарий для управления природными рисками при мореплавании автономных судов в Арктике при изменении климата / И. А. Сикарев, В. М. Абрамов, К. С. Простакевич [и др.] // Проблемы информационной безопасности. Компьютерные системы. – 2024. – № 1(58). – С. 110-120. – DOI 10.48612/jisp/v28t-z3kr-nrn2. – EDN RUESZV.
5. Priorities and challenges of the state policy of the Russian Federation in the Arctic science / G. Gogoberidze, V. Abramov, E. Rumyantseva [et al.] // 17th international multidisciplinary scientific geoconference SGEM 2017, Albena, Bulgaria, 29 June – 05 July 2017. Vol. 17. – Albena, Bulgaria, 2017. – P. 721-726. – DOI 10.5593/sgem2017/52/S20.092. – EDN ETLPDM.
6. Marine economic potential assessment for environmental management in the Russian Arctic and subarctic coastal regions / G. Gogoberidze, V. M. Abramov, L. N. Karlin [et al.] // 14th international multidisciplinary scientific geoconference SGEM 2014: Conference Proceedings, Albena, 17–26 June 2014. Vol. 3. – Sofia, 2014. – P. 253-260. – DOI 10.5593/SGEM2014/B53/S21.034. – EDN LDAIAX.6.
7. Web-based tools for natural risk management while large environmental projects / E. P. Istomin, V. M. Abramov, O. M. Lepeshkin [et al.] // 19th International Multidisciplinary Scientific GeoConference SGEM 2019: Conference proceedings. ENVIRONMENTAL ECONOMICS, Albena, 30 June – 06 July 2019. Vol. 19. – Sophia, 2019. – P. 953-960. – DOI 10.5593/sgem2019/5.3/S21.120. – EDN BHHPYQ.
8. Clean technologies development strategy for the national black carbon controlling system in the Russian Arctic / V. M. Abramov, G. G. Gogoberidze, L. N. Karlin [et al.] // 14th international multidisciplinary scientific geoconference SGEM 2014, Albena, Bulgaria, 17–26 June 2014. Vol. 2. – Sofia, 2014. – P. 313-320. – DOI 10.5593/SGEM2014/B42/S19.041. – EDN YONXKJ.
9. Abramov, V. M. Innovative geoinformation technologies within management of natural risks in Venezuela / V. M. Abramov, E. P. Istomin, J. A. Garcia // 18th International Multidisciplinary Scientific GeoConference SGEM 2018: Conference proceedings, Albena, Bulgaria, 02–08 July 2018. Vol. 18. – Albena, Bulgaria, 2018. – P. 261-268. – DOI 10.5593/sgem2018/2.2/S08.033. – EDN XKNRQX.
10. Abramov, V. M. Decadal variability of particulate matter in Moscow megacity air / V. M. Abramov, N. Popov, R. I. Bachiev // 16th International Multidisciplinary Scientific GeoConference SGEM 2016: Conference Proceedings, Albena, Bulgaria, 30 June – 06 July 2016. Vol. 2. – Albena, Bulgaria. – P. 323-330. – DOI 10.5593/SGEM2016/B42/S19.042. – EDN IHXCLN.
11. К вопросу о стратегии создания национальной системы контроля черного углерода в российской Арктике / Л. Н. Карлин, В. М. Абрамов, Г. Г. Гогоберидзе [и др.] // Ученые записки Российского государственного гидрометеорологического университета. – 2014. – № 36. – С. 67-73. – EDN TQUBAN.
12. Digital learning technologies within geo-information management / I. A. Sikarev, S. I. Lukyanov, N. Popov, [et al.] // E3S Web of Conferences, Chelyabinsk, 17–19 February 2021. – Chelyabinsk, 2021. – P. 01004. – DOI 10.1051/e3sconf/202125801004. – EDN GWVYAN.

Сикарев И.А.[1], Простакевич К.С.[1], Абрамова А.Л.[1], Абрамов В.М.[2]
[1]Российский государственный гидрометеорологический университет, Санкт-Петербург
[2]Государственный университет морского и речного флота имени адмирала С.О. Макарова, Санкт-Петербург

ЦИФРОВИЗАЦИЯ ДОВЕРЕННОГО ОБМЕНА ДАННЫМИ В КОНТЕКСТЕ КИБЕРБЕЗОПАСНОСТИ ИНТЕГРИРОВАННЫХ НАВИГАЦИОННЫХ СИСТЕМ В ЗОНЕ СЕВЕРНОГО МОРСКОГО ПУТИ

Проблема доверенного обмена данными (ДОД) в контексте кибербезопасности интегрированных навигационных систем (ИНС) в зоне Северного морского пути (СМП) является важной и актуальной [1 - 4] В настоящее время существуют достаточно жесткие нормативные требования Российской Федерации (РФ) к раскрытию данных об судне, его грузе и навигационных параметрах для допуска иностранного судна в зону Северного морского пути (СМП). При этом существует определенный правовой вакуум в сфере кибербезопасности навигации в зоне СМП. Существующие международные и национальные нормативные документы не регулируют допустимый предел доступа прибрежного государства к информации, хранящейся в судовых информационных системах. Целью исследования является разработка концептуальных решений проблемы ДОД в контексте информационной безопасности (ИБ) ИНС.

В ходе исследования использовались следующие методы и технологии:

--теория принятия решений в условиях неопределенности,

-Форсайт-технологии

- веб-технологии,

- технологии управления информационной безопасностью систем управления.

Перейдем к результатам исследования. С помощью системного анализа предметной области установлено, что традиционная система автоматической идентификационной системы (АИС) предполагает открытый обмен данными между судами. В зоне СМП на иностранного перевозчика накладывается дополнительное обременение – обязательное использование Российской системы дальней идентификации (РСДИ).

Если АИС в основном ограничивается обезличенным треком и скоростью, то РСДИ интегрируется с государственным порталом мониторинга, требуя более детального среза информации. С точки зрения ИБ ИНС, возникает риск принудительного «шлюзования» (NMEA-to-WAN). Подключая транспондеры к российской информационной среде, зарубежный капитан теряет контроль над целостностью навигационных данных. Появляется возможный вектор атаки «человек посередине» (Man-in-the-Middle), когда вмешательство в работу систем позиционирования может быть интерпретировано ИНС как легитимная корректировка от береговых служб.

Наиболее острой остается коллизия разрешительного порядка, закрепленного в ст. 79 Кодекса торгового мореплавания РФ и Правилах плавания в акватории СМП (Приказ Минтранса № 296). Для допуска к транзиту судовладелец через портал «Северный морской путь» обязан указать не просто технические характеристики судна, но и коммерческие параметры: порт назначения, наименование грузоотправителя и грузополучателя, тип и объем груза. С точки зрения безопасности, это требование создает так называемый «черный ящик» прибрежного государства. В нормативной логике РФ это требование трактуется как мера предотвращения загрязнения (ст. 234 UNCLOS). Однако для иностранной компании, указанное выше требование может рассматриваться как потенциальная угроза ИБ ИНС. Известно, что в 2024-2025 годах зафиксированы прецеденты, когда страхование грузов усложнялось из-за нарушения политик конфиденциальности компаний при вынужденном раскрытии информации о конечных получателях в Азиатско-Тихоокеанском регионе.

~~Статья 234 Конвенции ООН по морскому праву (UNCLOS) дает прибрежному государству право принимать законы по предотвращению загрязнения в ледовых условиях. Россия трактует это право расширительно, постулируя недискриминационность Правил плавания в акватории СМП, которые фактически устанавливают разрешительный, а не уведомительный режим прохода. Правовая позиция США и ряда морских держав заключается в том, что проливы Карского моря (Вилькицкого и др.) попадают под режим транзитного прохода, что исключает обязательную лоцманскую ледовую проводку и предварительное разрешение. Однако на практике технологическое доминирование России в ледокольном обеспечении, спутниковом мониторинге ледовой обстановки и услугах связи Iridium/RU-Link создает ситуацию вынужденной технологической зависимости.~~

Иностранное судно, следующее по СМП, будет вынужденно ~~интегрирует~~ интегрировать свои навигационные системы с береговыми серверами Атомфлота и Администрации СМП. В результате может возникнуть ~~Возникает~~ критическая уязвимость типа «~~слабого~~ слабое ~~звена~~звено» (Weaklest Link). Если ~~корпоративная~~ информационная сеть судна защищена последними версиями межсетевых экранов, то сегмент, отвечающий за связь с российским Штабом ледовых операций, часто функционирует на устаревших протоколах или требует отключения штатных средств защиты для корректной установки VPN-соединения. Это создает окно для спуффинга навигационных сигналов. Кроме того,~~,~~ получив контроль над каналом связи с берегом, злоумышленник может передать команду на корректировку курса, замаскированную под ледовый прогноз.

~~Действующие нормы, в частности рекомендации ИМО по кибербезопасности (Циркуляр MSC-FAL.1/Circ.3), фокусируются на автономной защите судна от атак вредоносного ПО и внутренних угроз экипажа. Они не рассматривают прибрежное государство в качестве потенциального актора угрозы для информационной безопасности коммерческого судна.~~

Анализ Полярного кодекса показывает, что он практически игнорирует коллизию цифровых юрисдикций. В результате судовладельцы оказываются в юридически конфликтных условиях, когда соблюдение нормативного акта Российской Федерации (РФ) о необходимости передачи данных о судне может вести к нарушению корпоративных, санкционных комплаенс-процедур и политик кибербезопасности, и наоборот.

Разрешение отмеченной выше правовой коллизии ~~видится~~ предлагается искать не в политическом компромиссе, а во внедрении криптографических технологий «нулевого разглашения (Zero-Knowledge Proofs, ZKP). В результате выполненного исследования предлагается использование архитектуры «Доверенного навигационного коридора» (Trusted Navigation Corridor, TNC), которая базируется на основе создания протокола, ~~в~~в рамках которого иностранное судно подтверждает выполнение требований безопасности (например, категорию ледового усиления Arc4/Arc7, наличие спасательных средств и т.д.) без раскрытия коммерчески чувствительной информации о грузе и владельце.

Технически это реализуемо через внедрение в ИНС судна аппаратного модуля безопасности, генерирующего криптографические доказательства. Администрация СМП получает верифицированный ответ «свойства судна соответствует допуску», не видя при этом данные о названии компании-фрахтователя или пункте назначения груза, если это не влияет на экологическую безопасность. Такой подход снимет правовую коллизию ст. 234 UNCLOS, сохранив суверенный контроль России над безопасностью судоходства и одновременно гарантировав судовладельцам иммунитет коммерческой тайны и защищенный периметр ИНС от внешнего цифрового вмешательства.

Результаты исследования обладают значительной научной новизной и могут быть использованы в различных направлениях [5 - 11], включая образовательные программы [12].

Платформа https://www.researchgate.net/profile/Valery_Abramov2/. использована в качестве инструмента коммуникации.

Список литературы:

1. Автоматизация деятельности водного транспорта с использованием факторного подхода к управлению рисками / И. А. Сикарев, А. Л. Абрамова, К. С. Простакевич [и др.] // Проблемы информационной безопасности. Компьютерные системы. – 2026. – № 1(68). – С. 168-175. – DOI 10.66424/2071-8217-2026-1-12. – EDN STKTVM.
2. Инфокоммуникационный инструментарий для управления природными рисками при мореплавании автономных судов в Арктике при изменении климата / И. А. Сикарев, В. М. Абрамов, К. С. Простакевич [и др.] // Проблемы информационной безопасности. Компьютерные системы. – 2024. – № 1(58). – С. 110-120. – DOI 10.48612/jisp/v28t-z3kr-nrn2. – EDN RUESZV.
3. Сикарев, И. А. Аспекты разработки и дальнейшие перспективы программы автоматической обработки спутниковых архивов гидрохимических данных на языке программирования Python / И. А. Сикарев, А. И. Честнов, В. М. Абрамов // Проблемы информационной безопасности. Компьютерные системы. – 2022. – № 4(52). – С. 101-109. – DOI 10.48612/jisp/7747-zanp-m6dr. – EDN NATOGA.
4. Цифровизация защиты информации в телекоммуникационных подсистемах полуавтономных надводных судов с внешним экипажем / И. А. Сикарев, В. М. Абрамов, К. С. Простакевич [и др.] // Методы и технические средства обеспечения безопасности информации. – 2024. – № 33. – С. 88-89. – EDN GMUWDH.
5. Priorities and challenges of the state policy of the Russian Federation in the Arctic science / G. Gogoberidze, V. Abramov, E. Rumyantseva [et al.] // 17th international multidisciplinary scientific geoconference SGEM 2017, Albena, Bulgaria, 29 June – 05 July 2017. Vol. 17. – Albena, Bulgaria, 2017. – P. 721-726. – DOI 10.5593/sgem2017/52/S20.092. – EDN ETLPDM.
6. Marine economic potential assessment for environmental management in the Russian Arctic and subarctic coastal regions / G. Gogoberidze, V. M. Abramov, L. N. Karlin [et al.] // 14th international multidisciplinary scientific geoconference SGEM 2014: Conference Proceedings, Albena, 17–26 June 2014. Vol. 3. – Sofia, 2014. – P. 253-260. – DOI 10.5593/SGEM2014/B53/S21.034. – EDN LDAIAX.6.
7. Web-based tools for natural risk management while large environmental projects / E. P. Istomin, V. M. Abramov, O. M. Lepeshkin [et al.] // 19th International Multidisciplinary Scientific GeoConference SGEM 2019: Conference proceedings. ENVIRONMENTAL ECONOMICS, Albena, 30 June – 06 July 2019. Vol. 19. – Sophia, 2019. – P. 953-960. – DOI 10.5593/sgem2019/5.3/S21.120. – EDN BHHPYQ.
8. Clean technologies development strategy for the national black carbon controlling system in the Russian Arctic / V. M. Abramov, G. G. Gogoberidze, L. N. Karlin [et al.] // 14th international multidisciplinary scientific geoconference SGEM 2014, Albena, Bulgaria, 17–26 June 2014. Vol. 2. – Sofia, 2014. – P. 313-320. – DOI 10.5593/SGEM2014/B42/S19.041. – EDN YONXKJ.
9. Abramov, V. M. Innovative geoinformation technologies within management of natural risks in Venezuela / V. M. Abramov, E. P. Istomin, J. A. Garcia // 18th International Multidisciplinary Scientific GeoConference SGEM 2018: Conference proceedings, Albena, Bulgaria, 02–08 July 2018. Vol. 18. – Albena, Bulgaria, 2018. – P. 261-268. – DOI 10.5593/sgem2018/2.2/S08.033. – EDN XKNRQX.
10. Abramov, V. M. Decadal variability of particulate matter in Moscow megacity air / V. M. Abramov, N. Popov, R. I. Bachiev // 16th International Multidisciplinary Scientific GeoConference SGEM 2016: Conference Proceedings, Albena, Bulgaria, 30 June – 06 July 2016. Vol. 2. – Albena, Bulgaria. – P. 323-330. – DOI 10.5593/SGEM2016/B42/S19.042. – EDN IHXCLN.
11. К вопросу о стратегии создания национальной системы контроля черного углерода в российской Арктике / Л. Н. Карлин, В. М. Абрамов, Г. Г. Гогоберидзе [и др.] //

Ученые записки Российского государственного гидрометеорологического университета. – 2014. – № 36. – С. 67-73. – EDN TQUBAN.
12. Digital learning technologies within geo-information management / I. A. Sikarev, S. I. Lukyanov, N. Popov, [et al.] // E3S Web of Conferences, Chelyabinsk, 17–19 February 2021. – Chelyabinsk, 2021. – P. 01004. – DOI 10.1051/e3sconf/202125801004. – EDN GWVYAN.

Синюков М.Д.
Федеральное учебно-методическое объединение в системе высшего образования по УГСН 10.00.00 «Информационная безопасность»

АНАЛИЗ АНОМАЛИЙ ФИЗИЧЕСКОГО УРОВНЯ LORA-СИГНАЛОВ С ЦЕЛЬЮ ВЫЯВЛЕНИЯ НЕЗАРЕГИСТРИРОВАННЫХ ПЕРЕДАТЧИКОВ

Аннотация. *Представлен гибридный метод анализа аномалий физического уровня LoRa-сигналов для выявления несанкционированных передатчиков. Разработана 14-мерная модель признакового пространства на основе CFO-профиля, спектральных и структурных характеристик пакетов. Предложен метод активного зондирования Probe-LoRa для обнаружения устройств, не участвующих в легитимном трафике, с подробным анализом его ограничений применительно к устройствам LoRaWAN Class A/B/C. Для принятия решения применён MAP-критерий на основе логистической регрессии (стекинг). Экспериментальная верификация на стенде USRP B210 (8 устройств, 3200 передач) показала AUC-ROC = 0,95, F1 = 0,93 при задержке обнаружения ≤ 340 мс. Явно обозначены систематические ограничения: малый датасет и неприменимость энтропийного признака к зашифрованным LoRaWAN-сетям.*

Ключевые слова: *LoRa, LoRaWAN, физический уровень, CFO, RF-фингерпринтинг, обнаружение аномалий, активное зондирование, Isolation Forest, CNN, SDR.*

1. Введение

Технология LoRa и протокол LoRaWAN широко применяются в инфраструктуре Интернета вещей – умных городах, промышленном мониторинге, сельском хозяйстве [1]. Работа в безлицензионных диапазонах (433/868/915 МГц) при уровне сигнала −120…−140 дБм затрудняет обнаружение нестандартных передатчиков. Существующие методы носят преимущественно пассивный характер [4, 5] и не обнаруживают устройства в режиме ожидания. Направление RF-фингерпринтинга LoRa на базе DL активно развивается [6, 7], однако большинство работ сосредоточено на аутентификации известных устройств. Настоящая работа предлагает гибридный метод, объединяющий пассивный анализ PHY-признаков с активным зондированием.

2. Модель признакового пространства физического уровня

Вектор признаков $f \in \mathbb{R}^{14}$ формируется на основе трёх групп параметров. Группа A (частотно-временные признаки f_1–f_4): смещение несущей частоты CFO – f_1, его стандартное отклонение σ_{CFO} – f_2, скорость изменения частоты в чирпе dF/dt – f_3, девиация длительности символа ΔT_s – f_4. Группа B (энергетические и спектральные f_5–f_8): RSSI, SNR, эксцесс IQ-выборок κ, коэффициент спектральной симметрии K_{sym}. Группа C (структурные f_9–f_{14}): sync word, длина преамбулы, энтропия payload, CR, IPI и отклик на зондирование Δt_resp [6, 8].

Критическое ограничение признака f_{11} (энтропия): стандартный LoRaWAN применяет AES-128 CTR, поэтому H(X) → 8 бит/байт как для легитимных, так и для аномальных пакетов. Признак информативен только для незашифрованных сетей. В эксперименте f_{11} исключён; эффективный вектор – 13 признаков.

Критерий статистической различимости устройств по CFO на основе отношения Фишера:

$$F_{sep} = \frac{(\mu i - \mu j)^2}{\sigma^2{}_i + \sigma^2{}_j} \geq F_{thr} = z^2{}_{1-\alpha/2} \cdot 2 \approx 7{,}68 \ (\alpha = 0{,}05) \quad (1)$$

Условие надёжного разделения: $|\mu_i - \mu_j| > 2{,}77 \cdot \sqrt{(\sigma_i^2 + \sigma_j^2)}$. При σ ≈ 100 Гц требуется $|\Delta\mu| > 390$ Гц. Устройства одной модели (Dragino LHT65: $F_{\mathrm{sep}} = 0{,}14$) оказались неразделимы, что подтверждает необходимость многопризнакового подхода. Ограничение: при ΔT = 20°C температурный дрейф CFO достигает 870–1740 Гц, сопоставимого с Δμ.

3. Метод активного зондирования Probe-LoRa

Метод направлен на стимуляцию ответной реакции устройств в режиме ожидания. Ключевое архитектурное ограничение [1]: устройства LoRaWAN Class A принимают downlink исключительно после собственной передачи – метод к ним неприменим. Class C (постоянный приём) и Class B (ping slots) являются целевыми классами.

Зондирующий фрейм использует тип Proprietary (MHDR.MType = 0b101) с криптографическим challenge для защиты от replay-атак [12]. Признак наличия ответа, его задержка Δt_{resp} и параметры модуляции (SF_{resp}, SW_{resp}) формируют признак f_{14} и критерий D_{probe}.

Экспериментально наблюдённые профили отклика трёх устройств-нарушителей стенда:

Таблица 1 – Профили отклика на зондирование (P_{resp}– эмпирические частоты)

Устройство (стенд)	$\mathbf{\Delta t_{resp}}$	SF_{resp}	SW_{resp}	P_{resp} **(n/N)**
Проприет. протокол #1 (N=30)	15–120 мс	SF7	0xFF	0,43 (13/30)
Тест. кан. управления #2 (N=30)	20–80 мс	SF9	0x12	0,67 (20/30)
Цифровой ретранслятор #3 (N=20)	80–250 мс	$= SF_{\text{p}}$	$= SW_{\text{p}}$	0,90 (18/20)

4. Ансамблевый классификатор и MAP-критерий

Применяется трёхкомпонентный ансамбль. Isolation Forest [13] выявляет аномалии в 13-мерном пространстве признаков. CNN (Conv2D×3 → FC256 → FC4) классифицирует STFT-спектрограммы преамбулы (64×64) [7] по четырём классам: соответствие PHY-профилю базы (C0), нестандартные параметры модуляции (C1), аномальный временной профиль (C2), реакция на зондирование (C3).

Итоговое решение – MAP-критерий на основе логистической регрессии второго уровня (стекинг):

$$P(C = 1|\cdot) = \sigma(w_0 + w_1 \cdot s_{IF} + w_2 \cdot P(C \neq 1|CNN) + w_3 \cdot D_{probe}) \quad (2)$$

Веса w_0–w_3 оценены MLE на обучающей выборке (5-кратная кросс-валидация), порог δ выбирается по ROC-кривой. Параметры байесовского обновления θ_P = 0,85 и τ_confirm = 3 также получены по ROC на данном стенде.

5. Экспериментальная верификация и результаты

Стенд: USRP B210, 5 легитимных устройств (Dragino LHT65 ×2, RAK7204, Pycom LoPy4 ×2), 3 устройства-нарушителя. 3200 передач (1800 легитимных / 1400 аномальных), несущая 868 МГц, BW = 125 кГц, SF ∈ {7,9,12} [14].

Таблица 2 – Сравнение методов (5-кратная кросс-валидация, лаб. стенд, 8 устройств)

Метод	**Prec.**	**Rec.**	**F1**	**AUC**	**Задержка**
Isolation Forest (пассивный)	0,87	0,81	0,84	0,89	< 8 мс
CNN / STFT (пассивный)	0,91	0,89	0,90	0,93	42 мс
Probe-LoRa (активный)	0,96	0,78	0,86	0,87	340 мс
Гибридный (стекинг, MAP)	**0,94**	**0,93**	**0,93**	**0,95**	**340 мс**

Примечание: AUC-ROC = 0,95 – верхняя оценка для лабораторных условий. Ожидаемое качество на незнакомых устройствах – AUC-ROC ≈ 0,70. Probe-LoRa повысил Recall для отвечающих устройств с 0,73 до 0,96, не изменив показатели для Class A.

6. Заключение

Предложен гибридный метод выявления несанкционированных LoRa-передатчиков, объединяющий пассивный анализ 13 PHY-признаков с активным зондированием. Теоретически обоснован CFO-критерий различимости (F_{sep}, F_{thr} = 7,68). Сформулированы строгие ограничения: Probe-LoRa неприменим к Class A; энтропия payload неинформативна при AES-128; AUC-ROC 0,95 является лабораторной оценкой. Направления развития: open-set датасет (> 50 устройств), анализ устойчивости к температурному дрейфу CFO, методы обнаружения при полном PHY-мимикрировании.

Список литературы:

1. LoRa Alliance. LoRaWAN Specification v1.0.4. LoRa Alliance Technical Committee, 2020.
2. Derache G. et al. LoRaWAN Attack in Military Use Case // arXiv:2412.18447. 2024.
3. Šabić J. et al. Practical Realization of Reactive Jamming Attack on LoRaWAN // Sensors. 2025. Vol.25. No.8. DOI: 10.3390/s25082383.
4. Senol N.S. et al. Identifying Tampered RF Transmissions in LoRa Using ML // Sensors. 2024. Vol.24. No.20. DOI: 10.3390/s24206611.
5. Machine Learning Models for LoRaWAN IoT Anomaly Detection // Proc. IEEE ICCWS. 2022. DOI: 10.1109/ICCWS56753.2022.9923439.
6. Ahmed A. et al. A Survey on Deep Learning-Based LoRa RF Fingerprinting // Sensors. 2024. Vol.24. No.13. DOI: 10.3390/s24134411.
7. Mahajan A., Burleson W. Watermarking and Anomaly Detection for LoRa RFFI // arXiv:2509.15170. 2025.
8. Vangelista L. Frequency Shift Chirp Modulation: The LoRa Modulation // IEEE SPL. 2017. Vol.24. No.12. P. 1818–1821.
9. Robyns P. et al. Physical-Layer Fingerprinting of LoRa // Proc. ACM WiSec. 2017. DOI: 10.1145/3098243.3098267.
10. Кобзарь А.И. Прикладная математическая статистика. М.: ФИЗМАТЛИТ, 2006.
11. Shannon C.E. A Mathematical Theory of Communication // Bell Syst. Tech. J. 1948. Vol.27.
12. Krawczyk H. et al. HMAC: Keyed-Hashing for Message Authentication. RFC 2104. 1997.
13. Liu F.T. et al. Isolation Forest // Proc. IEEE ICDM. 2008. P. 413–422.
14. Xaouri J. et al. SDR-LoRa: Open-Source LoRa on SDR // Computer Networks. 2024. DOI: 10.1016/j.comnet.2024.110126.

Соловьев Д.Ю., Зайцев В.В.

Военно-космическая академия имени А.Ф. Можайского, Санкт-Петербург

ПЕРСПЕКТИВЫ ПРИМЕНЕНИЯ ПАМЯТИ С НЕОСОЗНАВАЕМЫМ ДОСТУПОМ В ИНТЕРЕСАХ ЗАЩИТЫ ИНФОРМАЦИИ

Безопасность хранения данных – одно из ключевых требований в эпоху облачных вычислений, конфиденциального искусственного интеллекта и распределенных систем. Традиционные методы шифрования надёжно защищают содержимое данных, но оставляют уязвимым критически важный элемент – паттерны доступа. Злоумышленник, наблюдающий за обращениями к хранилищу на блочном уровне, промахами кэширования

или сетевым трафиком, способен восстановить чувствительную семантику: какие записи просматривает пользователь, структуру базы данных или поисковые запросы.

Исключить данную уязвимость способна память с неосознаваемым доступом последовательности логических обращений к данным в последовательность физических операций, статистически неотличимых друг от друга с точки зрения внешнего наблюдателя. Это достигается за счёт введения избыточных операций, случайного перераспределения данных и использования фиктивных обращений, маскирующих реальные действия пользователя.

Большинство практических реализаций использую схему Path ORAM [2]. При выполнении любой операции чтения или записи система загружает из серверного хранилища в локальную память пользователя все элементы, расположенные на пути от корня до листа его древовидной структуры. Данная структура состоит из узлов фиксированного размера. После извлечения или обновления требуемого блока производится перераспределение данных по новому случайному пути, дополнение фиктивными блоками и перезапись всего пути обратно в хранилище. В результате внешний наблюдатель видит только однородный случайный трафик, лишенный какой-либо корреляции с реальными действиями пользователя.

С теоретической точки зрения ORAM обладает строгой моделью безопасности, основанной на понятии неразличимости последовательностей обращений. Пусть последовательность логических обращений к памяти задается как

$$\mathrm{A} = (a_1, a_2, \dots, a_m),$$

где a_i – операция чтения или записи блока данных. В ORAM данная последовательность преобразуется в последовательность физических обращений

$$\tilde{A} = (\tilde{a}_1, \tilde{a}_2, \dots, \tilde{a}_t),$$

где t > m, а сами обращения не раскрывают информации о реальных действиях пользователя.

Требование неразличимости формализуется следующим образом: для любых двух логических последовательностей обращений A и $\tilde{A}$ одинаковой длины распределения соответствующих им физических последовательностей должны быть вычислительно неотличимы, то есть

$$\left| P\left(D\left(\tilde{A}\right)\right) - P\left(D\left(\tilde{A}'\right)\right)\right| < \varepsilon(N),$$

где D – произвольный вероятностный алгоритм-наблюдатель, а $\varepsilon(N)$ – пренебрежимо малая функция [1].

Важным аспектом является наличие фундаментального ограничения на эффективность ORAM. Обозначим через $C(N)$ амортизированную стоимость одного логического обращения, равную среднему числу физических операций, приходящихся на одну операцию доступа. Тогда для ORAM доказано существование нижней границы [3]

$$C(N) \geq \Omega(logN),$$

где N — число логических блоков в памяти.

Данная оценка означает, что при увеличении объема хранилища минимально необходимое число физических операций на одно логическое обращение растёт не медленнее логарифмической функции от размера памяти. В частности, если в традиционной системе доступ к одному блоку требует одной операции, то в ORAM для обеспечения неразличимости обращений необходимо выполнение не менее чем порядка $logN$ операций.

Логарифмическая зависимость накладных расходов обусловлена необходимостью сокрытия выбора одного блока среди множества из N возможных. Для обеспечения неразличимости система должна формировать такую последовательность обращений, которая не позволяет наблюдателю определить, какой именно элемент был запрошен. Объем дополнительной информации, необходимый для маскировки выбора, растёт логарифмически от мощности множества возможных обращений, что приводит

к соответствующей оценке сложности. Это позволяет рассматривать ORAM как технологию с заранее известным и строго обоснованным компромиссом между безопасностью и производительностью.

Перспективность применения ORAM существенно возрастает в контексте развития облачных вычислений. Современные модели хранения предполагают размещение данных на стороне внешнего провайдера. В таких условиях даже при отсутствии прямого доступа к данным анализ паттернов обращения позволяет выявлять структуру информации и поведенческие особенности пользователей. Применение ORAM позволяет устранить данный канал утечки, обеспечивая криптографически строгую защиту на уровне модели угроз.

Также направлением развития является использование ORAM в системах конфиденциальных вычислений, где требуется защита не только данных, но и процесса их обработки. Особую актуальность технология приобретает в задачах искусственного интеллекта, в которых информация о последовательности обращений к данным может раскрывать характеристики обучающих выборок и поведение пользователей.

Применение памяти с неосознаваемым доступом представляет собой перспективное направление развития средств защиты информации, обеспечивающее устранение фундаментального класса утечек, связанных с анализом паттернов обращения к данным.

Список литературы:

1. Oded Goldreich, Rafal Ostrovsky. Software Protection and Simulation on Oblivious RAMs [Электронный ресурс]. – URL: https://dl.acm.org/doi/ 10.1145/233551.233553 (дата обращения 03.04.2026).
2. Emil Stefanov, Marten van Dijk, Elaine Shi. Path ORAM: An Extremely Simple Oblivious RAM Protocol [Электронный ресурс]. – URL: https://eprint.iacr.org/2013/280 (дата обращения 03.04.2026).
3. Kasper Green Larsen, Jesper Buus Nielsen. Yes, There is an Oblivious RAM Lower Bound. [Электронный ресурс]. – URL: https://eprint.iacr.org/2018/423 (дата обращения 06.04.2026).

Шелухин О.И., Раковский Д.И.

Московский технический университет связи и информатики (МТУСИ), Москва

КАСКАДНАЯ МОДЕЛЬ УЧЕТА РЕДКИХ АНОМАЛЬНЫХ СОБЫТИЙ В ИНФОРМАЦИОННОЙ БЕЗОПАСНОСТИ

Публикация выполнена в рамках гранта на реализацию отраслевой научно-педагогической школы МТУСИ "Современные технологии исследования аномалий в информационной безопасности" по проекту "Обнаружение и прогнозирование редких аномальных событий для обеспечения информационной безопасности" (Пр. 93-х от 25.04.2025).

Под термином «редкие аномальные события» (*РАС или RAE - Rare Abnormal Event)* понимают события в определенной области, которые не происходят часто или регулярно, но имеют особое значение, или играют важную роль в системе, и их обычно трудно предсказать» [1]. В сфере информационной безопасности, при разработке систем обнаружения вторжений (СОВ), проблема обнаружения и классификация RAE заключается в дисбалансе количества данных, в которых миноритарный класс характеризуется «малым»

(<<10%) числом записей, в то время как мажоритарный класс – «большим» (подавляющим) числом записей, что характерно для редких событий.

Наряду с малым количеством данных, характерным для редких событий, весьма остро стоит проблема учета свойства многозначности имеющегося наборах данных, получаемых при обработке информации с детекторов сетевого трафика [2]. Источниками многозначных зависимостей являются как злонамеренные воздействия на компьютерную сеть (систему), так и коллизии, возникающие вследствие некорректной предобработки данных.

Вместе с предыдущими исследованиями, это позволяет определить редкое аномальное событие как реализацию многозначной зависимости, несущей деструктивное воздействие в процессе функционирования компьютерной системы (КС). С этой целью из всех событий, связанных с компьютерной системой *D*, необходимо отобрать те события, которые являются многозначными. После этого необходимо отобрать события, которые отвечают категории редкости, либо по частотному признаку, либо вероятностному. Среди отобранных событий также необходимо выделить те, которые несут некоторое деструктивное воздействие на компьютерную систему.

Пусть задано универсальное множество событий *D*, наблюдаемых в процессе функционирования КС включающее следующие подмножества: $D_M \subset D$ - подмножество многозначных событий, для которых выполняется условие: $\forall d \in D_M : \left|L(d)\right| \geq 2$, где $L(d) \subseteq \mathsf{L}$ - множество классовых меток, ассоциированных с событием *d*, L - алфавит уникальных типов компьютерных атак (КА). $D_A \subset D$ - подмножество аномальных (деструктивных) событий, семантически определяемое как события, нарушающие свойства доступности, целостности или конфиденциальности информации. $D_R \subset D$ - подмножество редких событий, определяемое через вероятностную меру P - $\forall d \in D_R : \mathsf{P}(d) < \varepsilon, \varepsilon \leq 0{,}01$, где ε — порог редкости [3], согласно которому событие относится к категории наименее вероятных («редких событий»), если вероятность его возникновения равна или менее 0,01.

С учетом изложенного, ***редкое аномальное событие*** может быть формально определено как

$$RAE = D_A \cap D_M \cap D_R . \quad (1)$$

Актуальной является задача интеграции информации о RAE в современные системы обнаружения или предотвращения вторжений.

Целью работы является разработка каскадной модели описания компьютерной системы, учитывающей наличие редких аномальных событий в многозначной трактовке.

Объединение трёх категорий событий в (1) является лишь малой частью диаграмм Эйлера-Венна обычно используемых для наглядного представления отношений между множествами для понимания и анализа логических связей. Поэтому необходимо рассмотреть все возможные комбинации объединения трех представленных подмножеств.

Комбинация $D_A \cap D_M \cap D_R$ характеризует ***редкие многозначные аномальные события.*** Как правило это сложные, ранее не встречавшиеся деструктивные воздействия, одновременно нарушающее несколько свойств информационной безопасности и комбинирующее два и более типов аномалий и/или КА.

Комбинация $D_A \cap D_M \cap \bar{D}_R$ характеризует ***частые многозначные аномальные события*** , *а* $\bar{D}_R$ - это множество всех «не редких (штатных)» событий. Это могут быть регулярно повторяющиеся комбинированные атаки или масштабные кампании, уже зафиксированные в исторических данных. Многозначность в записях при этом присутствует, но ее вероятностная мера превышает порог редкости. Множество частых событий $\bar{D}_R$ определяется как $\bar{D}_R = \left\{ d \in D_R \middle| \mathsf{P}(d) > \varepsilon \right\}$, где ε — порог редкости.

Следовательно, для подмножества частых многозначных аномальных событий выполняется условие:

$$\forall d \in D_A \cap D_M \cap \bar{D}_R : P(d) > \varepsilon . \quad (2)$$

Комбинация $D_A \cap \bar{D}_M \cap D_R$ характеризует ***редкие однозначные аномальные события,*** где $\bar{D}_M$ - множество однозначных событий. Как правило это единичная, статистически редкая атака, направленная на один вектор воздействия или эксплуатирующая изолированную уязвимость, а многозначные зависимости отсутствуют.

Комбинация $\bar{D}_A \cap D_M \cap D_R$ характеризует ***редкое многозначное штатное событие,*** где $\bar{D}_A$ - множество штатных событий**.** Это могут быть статистически редкие легитимные паттерны трафика, не несущие деструктивного воздействия.

Комбинация $D_A \cap \bar{D}_M \cap \bar{D}_R$ характеризует ***частое однозначное аномальное событие.*** Например, это могут быть типовые, хорошо изученные одиночные атаки, регулярно наблюдаемые в сети.

Комбинация $\bar{D}_A \cap D_M \cap \bar{D}_R$ характеризует ***частое многозначное штатное событие.*** В этом случае штатный сложный трафик, естественно порождающий многозначные зависимости, однако без нарушения свойств ИБ.

Комбинация $\bar{D}_A \cap \bar{D}_M \cap D_R$ характеризует ***редкое однозначное штатное*** *событие и* является полной противоположностью ***штатному нормальному трафику*** - $\bar{D}_A \cap \bar{D}_M \cap \bar{D}_R$**.**

На основании рассмотренных комбинаций предлагается каскадная модель обнаружения аномальных событий, представленная в таблице 1. Модель может быть интегрирована в эшелонированную структуру СОВ, состоящую из нескольких модулей: модуль реагирования на атаки нулевого дня (M_1), модуль реагирования на многозначные аномалии (M_2); модуль реагирования на однозначные аномалии (M_3); модуль сигнатурного анализа (M_4).

Таблица 1. Каскадная модель поведения КС, учитывающая RAE

№	Область	$D_A \subset D$	$D_R \subset D$	$D_M \subset D$	Тип события	Модуль
1	$D_A \cap D_M \cap D_R$	+	+	+	Редкое многозначное аномальное событие	**M_1**
2	$D_A \cap D_M \cap \bar{D}_R$	+	-	+	Частое многозначное аномальное событие	M_2
3	$D_A \cap \bar{D}_M \cap D_R$	+	+	-	Редкое однозначное аномальное событие	M_3
4	$\bar{D}_A \cap D_M \cap D_R$	-	+	+	Редкое многозначное штатное событие	**M_1**
5	$D_A \cap \bar{D}_M \cap \bar{D}_R$	+	-	-	Типовая аномалия	M_4
6	$\bar{D}_A \cap D_M \cap \bar{D}_R$	-	-	+	Штатный трафик, обладающий свойством многозначности	M_2
7	$\bar{D}_A \cap \bar{D}_M \cap D_R$	-	+	-	Редкое штатное событие	M_3
8	$\bar{D}_A \cap \bar{D}_M \cap \bar{D}_R$	-	-	-	Штатный трафик	M_4

Предложенная модель позволяет выполнять категоризацию событий до их отправки на специализированные модули реагирования на конкретные типы аномалий, что оптимизирует распределение вычислительных ресурсов на модулях.

Список литературы:

1. Шелухин О.И., Раковский Д.И. Редкие аномальные события. проблемы обнаружения и обработки // Наукоемкие Технологии В Космических Исследованиях Земли. 2025. Т. 17, № 4. С. 54–68. DOI: 10.36724/2409-5419-2025-17-4-54-68.

2. Шелухин О.И., Раковский Д.И. Многозначная классификация и прогнозирование. М.: НТИ «Горячая линия - Телеком». 296 с.
3. Shyalika C., Wickramarachchi R., Sheth A.P. A Comprehensive Survey on Rare Event Prediction // ACM Comput. Surv. 2025. Т. 57, № 3. С. 1–39. DOI: 10.1145/3699955.

Штеренберг С.И.
Санкт-Петербургский государственный университет телекоммуникаций им. проф. М.А. Бонч-Бруевича, Санкт-Петербург

СТЕГАНОГРАФИЧЕСКИЙ ПОДХОД К ЗАЩИТЕ ДЕЦЕНТРАЛИЗОВАННЫХ РАССИНХРОНИЗИРОВАННЫХ ПАКЕТНО-НЕЙРОСЕТЕВЫХ ПРОГРАММ

Актуальность исследования обусловлена необходимостью обеспечения киберустойчивости интеллектуальных систем обнаружения вторжений (СОВ) в условиях децентрализации и автономного функционирования их компонентов. Программные агенты (ПА), образующие основу пакетно-нейросетевых программ (ПНП), уязвимы для целенаправленных атак, особенно при переходе системы в децентрализованное и рассинхронизированное состояние. Существующие решения на основе ассимиляционной памяти и самоорганизующихся карт не обеспечивают комплексной защиты и не задают плана действий при каскадных сбоях и отказах в обслуживании [1,2]. Целью работы является разработка стеганографического подхода, объединяющего модель выявления угроз, алгоритм стеганографических операций и методику контейнеризации ПА для обеспечения скрытности, самовосстановления и защиты от отказов в обслуживании.

Модель выявления угроз нарушения информационной безопасности ПНП основана на детектировании признаков рассинхронизации децентрализованных агентов. Вводится логическая функция распознавания, инспирированная картами Карно:

$$\theta = \alpha\beta\bar{f} \vee \beta\overline{e \vee \bar{\alpha}\bar{\beta}}, \quad (1)$$

где α, β – активность и целостность ПА; e – признак централизованного управления; f – синхронизация.

Частные значения $\theta_1 = \theta(f=0)$ и $\theta_0 = \theta(f=1)$ определяют интервал существования агента PiRun. Закономерности формализуются трёхмерной матрицей T, комбинации элементов которой интерпретируются как состояния системы. Для количественной оценки угроз вводится коэффициент превосходства f_i и результирующая матрица r, итоговая оценка $H_i = \sum_{n=1}^{N} r_{n,i}$. Чем выше H_i, тем более критично состояние агента и необходимее переход к стеганографической защите.

Алгоритм стеганографических операций базируется на концепции «Безопасного сетевого червя» и использует Hid-код – скрытое вложение, встраиваемое непосредственно в исполняемые файлы. Стойкость стеганосистемы обеспечивается энтропийным подходом: условие $H(C) + D(P_c||P_s) = log\,|X|$ гарантирует неразличимость пустых и заполненных контейнеров при минимизации дивергенции Кульбака–Лейблера. Качество преобразования оценивается энтропией сообщения, а имитозащита – функцией $X(S,K_i)$, отвергающей имитонавязанные сообщения. Классовая политика внедрения Hid-кода включает три уровня: Класс I (пустоты, сжатие), Класс II (расширение заголовка, оверлеи), Класс III (расширение/создание секций, «подчистка»). Суммарная квадратичная ошибка SE(S) позволяет оценить скрытность. Практический алгоритм реализует полный цикл от разбора PE-файла до обратной сборки с полиморфным генератором, обеспечивающим уникальную бинарную сигнатуру каждой копии агента.

Методика контейнеризации гарантирует сохранение стеганопреобразований на всех этапах жизненного цикла ПА. Проанализированы угрозы: антипаттерны Dockerfile,

уязвимости базовых образов, дефекты конфигурации, проблемы журналирования. Методика включает шесть обязательных шагов: выбор минимального образа, проверка цепочки поставок (Sigstore/Cosign), генерация Dockerfile с multi-stage сборкой, минимизацией слоёв, оптимизацией зависимостей, ограничением прав (USER nonroot), шифрованием томов (LUKS), сканированием уязвимостей (Trivy/Snyk), а также применение AppArmor/Seccomp/SELinux, сетевых политик Kubernetes, Docker Secrets, фиксацию тэгов и запуск в режиме read-only с ротацией логов.

Экспериментальная верификация проводилась на специализированном стенде. Для анализа временных рядов использовались условие нормировки $1 \leq X_t \leq P_m, \sum_{i=1}^{3} a_i + b_i = r(k)$, коэффициенты парной и множественной корреляции. Исследованы два темпа «мыслительного» процесса YaVi: Темп 1 (деградация данных, невозможность упаковки в СУБД) и Темп 2 (кластеризация, восстановление структуры). Количественная оценка эффективности стеганографических операций представлена в табл. 1.

Таблица 1 – Влияние обфускации на сохранение ЦВЗ

Размер файла, кбайт	Длина ЦВЗ, бит	Тесты обфускации		Процент сохранения ЦВЗ	Примечание
		тест 1	тест 2		
542,1	4032	1	-	51%	запускается
		-	1	49%	не запускается
		2	1	36 %	не запускается
		1	2	32%	не запускается
802,2	6144	1	-	48%	запускается
		-	1	46%	не запускается
		2	1	37 %	не запускается
		1	-	34%	не запускается
208, 9	3240	1	-	52%	запускается
		-	1	39%	запускается
		2	1	36 %	не запускается
		1	-	32%	не запускается

Как видно из таблицы, процент сохранения вложенных ЦВЗ достигает 52% при однократной обфускации. При этом снижение скорости ПА находится в диапазоне 26–37%. Полученные результаты подтверждают эффективность предложенных решений по обеспечению живучести ПНП в деструктивной среде.

Таким образом, разработанный стеганографический подход, объединяющий модель идентификации угроз, алгоритм скрытого вложения и методику защищённой контейнеризации, позволяет обеспечить киберустойчивость интеллектуальных СОВ в условиях децентрализации и отказов в обслуживании.

4. МЕТОДЫ АНАЛИЗА БЕЗОПАСНОСТИ ПРОГРАММНОГО ОБЕСПЕЧЕНИЯ

Аношкин И.А., Пагуба Г.Ю., Кондачков Е.Д.
Санкт-Петербургский политехнический университет Петра Великого, Санкт-Петербург

ОЦЕНКА ДОСТИЖИМОСТИ УЯЗВИМЫХ ФРАГМЕНТОВ ПРОГРАММНОГО КОДА ПРИЛОЖЕНИЙ ДЛЯ ОС ANDROID

По состоянию на 2026 год количество смартфонов на базе ОС Android превышает 3,9 миллиардов и составляет около 73% мирового рынка [1] и такое широкое распространение обусловлено не только необходимостью обеспечения связи. Смартфон сегодня является персональным идентификационным устройством, кошельком и хранилищем документов и других персональных данных, в следствие чего его защищенность становится первоочередной задачей для производителя аппаратных и программных составляющих. Однако, подход, применяемый компанией производителем, не всегда применим для стороннего разработчика, приложение которого распространяется через, например, Google Play. Ошибки в подобных программах подвергают опасности большой массив персональной информации пользователя, так как создают возможность её компрометации злоумышленником. Так, согласно статистическим данным [2] большая часть приложений, распространяемых через Google Play, содержат в себе уязвимости высокого уровня критичности, что естественным образом приводит к стремительному росту количества атак на мобильные устройства. Таким образом, приоритетным направлением исследования сегодня является быстрый поиск и устранение уязвимостей в Android-приложениях.

В предыдущих работах авторов [3,4] представлены методы автоматизированного анализа программного обеспечения для ОС Android на основе методов статического и динамического анализа. Представленные методы могут быть использованы для поиска криптографических уязвимостей, уязвимостей сторонних программных компонентов и выявления хранения и передачи чувствительной информации в открытом виде. Представленные ранее методы обеспечивают достаточное покрытие функционала приложения для оценки его безопасности по каждой из категорий выделенных в OWASP Mobile Top 10 2024 [5].

Одним из наиболее явных недостатков представленных ранее методов является большое количество ложных срабатываний. Несмотря на важность поиска уязвимостей в программном обеспечении, собранная информация не будет полной и релевантной, пока не будет проведена оценка этих уязвимостей с точки зрения достижимости. Под достижимостью в данной работе понимается возможность использования уязвимого участка кода в ходе штатной работы программы. Разумно будет предположить, что недостижимые участки кода с меньшей вероятностью могут породить опасную последовательность действий, которые приводят к нежелательным последствиям. Для оценки достижимости фрагментов программного кода приложений для ОС Android предлагается использовать модифицированное сочетание прямого и обратного символьного исполнения [6]. Ключевой проблемой для использования классического символьного исполнения в контексте анализа достижимости фрагментов программного

кода приложений для ОС Android является использование сочетания нативного программного кода, написанного на языках программирования Си/C++ и Rust с программным кодом, написанным на языках программирования Java и Kotlin. Данную проблему предлагается решать за счет разделения символьных доменов для программного кода, написанного на различных языках и агрегации символьных ограничений перед их разрешением.

Предлагаемый метод имеет следующие основные этапы:

1. Выбор фрагмента программного кода, оценку достижимости которого необходимо произвести.
2. Определение фрагмента программного кода, с которого необходимо оценить достижимость уязвимого фрагмента. В качестве данного фрагмента может быть выбрана точка входа в нативную библиотеку приложения (метод JNI_onLoad) или метод onCreate класса Activity приложения.
3. Сбор символьных ограничений для нативного кода и для кода на языках программирования Java/Kotlin. Стоит отметить, что на данном этапе используется два разных экземпляра символьных решателей.
4. Определение междоменных ограничений. Под междоменными ограничениями понимаются символьные ограничения, описывающие аргументы функций нативного кода, вызываемые из программного кода на языках Java и Kotlin.
5. Нормализация междоменных символьных ограничений.
6. Агрегация символьных ограничений обоих доменов и междоменных ограничений.
7. Разрешение символьных ограничений.
8. Формирование вывода о достижимости уязвимого фрагмента программного кода из выбранного фрагмента.

Таким образом, в работе предложен метод оценки достижимости уязвимых фрагментов программного кода в контексте приложений для ОС Android. Предложенный метод позволит уменьшить количество ложных срабатываний методов поиска уязвимостей и ошибок в приложениях для ОС Android.

Список использованных источников:

1. Android Usage Statistics 2026 [Электронный ресурс]. URL: https://www.demandsage.com/android-statistics/. – (дата обращения: 10.05.2026).
2. Mobile Security Statistics in 2026: Trends, Threats & Insights for a Safer Future [Электронный ресурс]. URL: https://comparecheapssl.com/mobile-security-statistics-trends-threats-insights-for-a-safer-future/. – (дата обращения: 10.05.2026)
3. Е. Ю. Павленко, И. А. Аношкин. Автоматизированный анализ безопасности программного обеспечения для ОС Android / Павленко Е. Ю. // Проблемы информационной безопасности. компьютерные системы. – 2025. – №. 2. – С. 112-123.
4. Пагуба Г.Ю., Волков А.А., Самарин Н.Н. Применение алгоритма выравнивания по волновому фронту для поиска потенциально-уязвимых компонентов с открытым исходным кодом в приложениях для радиотелефонов на базе ОС Android // Радиотехника. 2026. Т. 90. № 2. С. 66−72. DOI: https://doi.org/10.18127/j00338486-202602-09.
5. OWASP Mobile Top 10 [Электронный ресурс]. URL: https://owasp.org/www-project-mobile-top-10/. – (дата обращения: 12.05.2026).
6. Самарин Н. Н, Шелухин О. И. Метод оценки достижимости фрагментов программного кода с применением символьного исполнения // Наукоемкие технологии в космических исследованиях Земли. 2024. Т. 16 № 4. С. 33 – 40. doi: 10.36724/2409-5419-2024-16-4-33-40

Волков А.А.[1], Пагуба Г.Ю.[1], Самарин Н.Н.[2]
[1] Санкт-Петербургский политехнический университет Петра Великого, Санкт-Петербург
[2] Ордена Трудового Красного Знамени федеральное государственное бюджетное образовательное учреждение высшего образования «Московский технический университет связи и информатики» (МТУСИ), Москва

СИНТЕЗ МОДИФИКАЦИЙ ПРОГРАММНОГО КОДА НА ОСНОВЕ ПРОМЕЖУТОЧНОГО ПРЕДСТАВЛЕНИЯ LLVM

Одной из ключевых проблем при автоматической модификации программного обеспечения является необходимость генерации архитектурно-корректного программного кода. На этом уровне изменения требуют учёта архитектурных особенностей процессора, соглашений о вызовах, организации стека, регистровой модели и точной структуры управляющего потока. При этом распознавание логических конструкций программы по набору инструкций представляет собой нетривиальную задачу, поскольку один и тот же фрагмент алгоритма может быть реализован различными последовательностями машинных инструкций, а границы функций, условий, циклов и вызовов не всегда удаётся определить однозначно.

Современные подходы к модификации бинарного кода на уровне ассемблера обладают рядом недостатков:

1. На уровне ассемблера отсутствует явное представление о логике программы, вследствие чего идентификация проверок, ветвлений, вычислительных операций и точек принятия решений требует значительных аналитических усилий.
2. Автоматическое выявление мест, в которых следует изменить логику программы, затруднено из-за отсутствия устойчивых признаков высокоуровневых конструкций.
3. Ассемблерный код жёстко привязан к конкретной архитектуре процессора и соглашениям о вызовах операционной системы, что ограничивает переносимость методов анализа и модификации.

В связи с указанными ограничениями использование промежуточного представления LLVM IR[1] представляет собой более удобный подход к анализу и модификации программного кода. LLVM IR обладает более высокой степенью абстракции и формализованной структурой, что позволяет выполнять преобразования на уровне функций, ветвлений, вызовов и возвращаемых значений. Это, в свою очередь, упрощает синтез модификаций и снижает сложность их автоматического применения.

Предлагаемый метод основан на использовании LLVM IR[2] как промежуточного представления, позволяющего перенести задачу модификации программного кода с уровня ассемблерных инструкций на более высокий и формализованный уровень абстракции. Исходным объектом анализа выступает бинарный модуль формата ELF, содержащий машинный код для различных архитектур. На первом этапе выполняется восстановление структуры бинарного файла, выделение функций и базовых блоков, после чего машинный код транслируется в промежуточное представление с последующей генерацией LLVM IR. Такой подход позволяет сохранить логическую структуру программы и сделать её пригодной для автоматической трансформации без доступа к исходным текстам.

Дальнейшая модификация кода осуществляется непосредственно на уровне LLVM IR за счёт применения преобразований, заданных с помощью S-выражений [3] и ориентированных на изменение поведения отдельных функций и участков управляющего потока. В рамках предлагаемого метода возможны подмена возвращаемых значений, изменение результатов сравнений, удаление или подмена вызовов функций, а также перенаправление ветвлений по заданной логике. В отличие от прямого редактирования ассемблера, такие преобразования описываются в более компактной и семантически

понятной форме, что снижает вероятность нарушения корректности исполнения и упрощает автоматизацию процесса модификации.
После внесения изменений модифицированное LLVM IR подвергается валидации и последующей компиляции обратно в исполняемый объектный код. Далее изменённый машинный код может быть внедрён непосредственно в оригинальный ELF-файл с сохранением остальной структуры бинарного файла. Таким образом, предлагаемый метод обеспечивает управляемую и архитектурно ориентированную модификацию программного кода, позволяя реализовывать точечные изменения поведения бинарных модулей без повторной сборки исходного программного обеспечения.

Список использованных источников:

1. LLVM Project. LLVM Language Reference Manual. [Электронный ресурс]. URL: https://llvm.org/docs/LangRef.html. – (дата обращения: 08.05.2026).
2. Lattner C., Adve V. LLVM: A Compilation Framework for Lifelong Program Analysis & Transformation // Proceedings of the International Symposium on Code Generation and Optimization (CGO). 2004. P. 75–86. – (дата обращения: 10.05.2026).
3. Real World OCaml. Data Serialization. [Электронный ресурс]. URL: https://dev.realworldocaml.org/data-serialization.html. – (дата обращения: 08.05.2026).

УДК 004.056

Гололобов Н.В.
Санкт-Петербургский политехнический университет Петра Великого, Санкт-Петербург

ФОРМАЛЬНАЯ МОДЕЛЬ ИСПОЛНЯЕМОГО КОДА ПРОГРАММНОГО ОБЕСПЕЧЕНИЯ НА ОСНОВЕ ГРАФА ИСПОЛНЕНИЯ ДЛЯ РЕШЕНИЯ ЗАДАЧИ БИНАРНОГО МОРФИНГА

Разработка формальной математической модели программного обеспечения является необходимым условием для строгого определения операций динамического бинарного морфинга и доказательства их корректности. Без формального аппарата, позволяющего однозначно описывать структуру исполняемого кода и отношения между его элементами, невозможно дать точное определение функциональной эквивалентности двух образцов кода и сформулировать доказуемые гарантии её сохранения при преобразованиях.

Анализ известных моделей программного обеспечения показал, что с точки зрения полноты математического описания программы известные модели не являются взаимоисключающими, а скорее дополняют друг друга [1-5]. Логическая и аксиоматическая модели лучше подходят для доказательства корректности, автоматная модель – для анализа поведения во времени, грамматическая – для описания структуры, функциональная – для представления вычисления как отображения, а информационно-структурная – для исследования потоков данных и организации вычислений. Для формального анализа программного обеспечения требуется сочетание нескольких моделей одновременно [6]. Например, синтаксическая структура программы может быть задана грамматически, её поведение – автоматно, а свойства корректности – логически.

Математическая модель программного обеспечения, разрабатываемая в контексте задачи динамического морфинга, должна удовлетворять ряду ключевых требований, определяемых как общими принципами математического моделирования, так и специфическими особенностями решаемой задачи. Определены следующие требования: полнота, формальность, операциональность, применимость к бинарному коду, вычислительная обозримость [7, 8].

в контексте текущих исследований, называется тройка:

=

)

где $ICFG(P)$ – межпроцедурный граф потока управления программы; к₀ – начальная конфигурация расположения кода, устанавливаемая загрузчиком; $\Gamma(\kappa_0)$ – множество управляющих ссылок в начальной конфигурации. Формальная модель полностью описывает статическую структуру программы, существенную для анализа операций морфинга [20].

Следует подчеркнуть, что формальная модель $\mathcal{M}(P)$ абстрагируется от конкретного бинарного представления инструкций и оперирует исключительно понятиями базовых блоков, переходов и адресов. Это позволяет рассматривать операции морфинга независимо от конкретной кодировки инструкций целевой архитектуры.

Динамический уровень формальной модели описывает фактическое исполнение программы как последовательность состояний вычислительного процесса. Данный уровень необходим для формального определения функциональной эквивалентности: две программы эквивалентны в том и только том случае, если их динамические характеристики совпадают.

Пусть дана программа P с формальной моделью $\mathcal{M}(P) = (ICFG(P), \kappa_0, \Gamma(\kappa_0))$. В результате операции морфинга получается новая конфигурация расположения кода κ_1, отличная от κ_0. При этом необходимо определить, при каком условии программа в к

о

Тогда конфигурацию расположения кода κ_0 и κ_1 программы P можно определить как

функционально эквивалентные (обозначение: $\kappa^0 \approx_P \kappa_1$) тогда и только тогда, когда для

любых входных данных $d \in D$, где D – множество допустимых входных данных

программы P, выполняется:

г =

у г)

р

а

ц

и

Разработанная формальная модель программного обеспечения и операций

динамического бинарного морфинга обладает высокой степенью общности и теоретической

строгостью. Тем не менее, она опирается на ряд упрощающих предположений, которые не

в полной мере отражают особенности реального программного обеспечения.

ф Наиболее чувствительным ограничением является предположение о статической и

полностью восстанавливаемой структуре CFG. Другим ограничением модели являются

динамически преобразовываемые программы, к которым относится

самомодифицирующийся код или механизм динамической загрузки кода

а Формальная модель определена в архитектурно-независимых терминах: понятия

базового блока, CFG, управляющих ссылок и наблюдаемого поведения применимы к любой

архитектуре. Тем не менее конкретные детали реализации – размеры инструкций,

кодирование операндов переходов, механизм таблиц переходов – существенно различаются

для архитектур x86-64, ARM64 и RISC-V. Дальнейшие исследования должны быть

направлены на сокращение количества ограничений модели и обеспечение ее практической

реализации поддержки архитектур ARM64 и RISC-V.

Список источников:

1. Chand D. R., Yadav S. B. Logical construction of software //Communications of the ACM. – 1980. – Т. 23. – №. 10. – С. 546-555.
2. Posnett D., Hindle A., Devanbu P. A simpler model of software readability //Proceedings of the 8th working conference on mining software repositories. – 2011. – С. 73-82.

3. Milicev D., Jovanovic Z. A finite state machine based format model of software pipelined loops with conditions //Progress in Computer Research. – 2001. – С. 51.
4. Arafeh B. R. A graph grammar model for concurrent and distributed software specification-in-large //Journal of Systems and Software. – 1995. – Т. 31. – №. 1. – С. 7-32.
5. Davis D. Modelled on software engineering: Flexible parametric models in the practice of architecture: дис. – RMIT University, 2024.
6. Wang Y. On the informatics laws and deductive semantics of software //IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews). – 2006. – Т. 36. – №. 2. – С. 161-171.
7. Heisel M., Santen T., Souquieres J. Toward a formal model of software components //International Conference on Formal Engineering Methods. – Berlin, Heidelberg : Springer Berlin Heidelberg, 2002. – С. 57-68.
8. Molan G. et al. Model for Quantitative Estimation of Functionality Influence on the Final Value of a Software Product //IEEE Access. – 2023. – Т. 11. – С. 115599-115616.

Грибков Н.А., Калинин М.О.
Санкт-Петербургский политехнический университет Петра Великого, Санкт-Петербург

СЕМАНТИЧЕСКИЙ КОМПОЗИЦИОННЫЙ АНАЛИЗ ПРОГРАММНОГО КОДА

В задачах анализа программного кода для информационной безопасности особое значение имеют методы, способные учитывать не только синтаксическую форму фрагментов, но и их семантику. Это связано с тем, что уязвимости, клоны кода и повторно используемые компоненты часто сохраняют функциональное сходство при существенных синтаксических преобразованиях. Кроме того, современные исследования показывают, что графовые представления, абстрактные синтаксические деревья и семантические промежуточные представления позволяют повысить качество поиска схожих фрагментов кода по сравнению с чисто синтаксическими методами.

Семантический композиционный анализ рассматривает смысл программного фрагмента как результат объединения смыслов его частей: инструкций, выражений, базовых блоков и функций. В отличие от сравнения строк или токенов на уровне синтаксиса, такой подход опирается на структуру выполнения и отношения между элементами кода, что важно при анализе двоичных программ, деобфусцированного кода и кода, подвергшегося оптимизациям компилятора.

В работе предлагается метод, основанный на автоматическом формировании контекстно-обогащенных графовых представлений программ, а также проводится оценка его эффективности. Ключевая идея метода состоит в том, что семантика фрагмента кода должна оцениваться не изолированно, а через композицию смыслов его частей: инструкций, базовых блоков, функций и межкомпонентных связей. Такой подход позволяет сопоставлять программные фрагменты даже при различиях в синтаксисе, компиляторных преобразованиях и изменениях в архитектуре целевой платформы.

Для реализации метода программный код представляется в виде формальной модели, включающей компоненты, зависимости, фрагменты кода и их контекстно-обогащенные описания. На основе этой модели строится унифицированное промежуточное представление, инвариантное к исходной форме программы и пригодное как для исходного текста, так и для бинарного кода. Для двоичных программ выполняется восстановление кода, построение абстрактного синтаксического дерева и дальнейшее расширение до атрибутного дерева с семантическими признаками.

Численная оценка семантической схожести фрагментов реализуется с помощью графовой нейронной сети с сиамской архитектурой, где представление кода отображается в вектора признаков. Сравнение векторов позволяет выделять как синтаксические, так и семантические клоны, включая слабо похожие фрагменты третьего и четвертого типа. Полученная метрика используется для сопоставления анализируемых фрагментов с репозиториями уязвимых и доверенных шаблонов.

Преимуществом семантического композиционного анализа является его устойчивость к переименованию переменных, перестройке выражений и другим локальным модификациям, которые затрудняют применение классических методов сопоставления. Это делает подход перспективным для задач поиска уязвимых фрагментов, обнаружения повторного использования библиотечного кода и анализа программных компонентов в условиях отсутствия исходных текстов. В более широком контексте подобные методы дополняют инструменты статического анализа и могут использоваться в системах поддержки при экспертном анализе безопасности.

Список литературы:

1. Gribkov N. A., Ovasapyan T. D., Moskvin D. A. Analysis of Decompiled Program Code Using Abstract Syntax Trees // Automatic Control and Computer Sciences. 2023. Vol. 57. No. 8. P. 958–967.
2. Gribkov N. A., Ovasapyan T. D., Moskvin D. A. Binary Code Preprocessing Method Based on Pseudocode Analysis for Binary Code Similarity Detection Task.
3. Song H.-J., Park S.-B., Park S. Y. Computation of Program Source Code Similarity by Composition of Parse Tree and Call Graph // Mathematical Problems in Engineering. 2015.
4. Xu X., Liu C., Feng Q. et al. Neural Network-based Graph Embedding for Cross-Platform Binary Code Similarity Detection // CCS. 2017.
5. Yang S., Cheng L., Zeng Y. et al. Asteria: Deep Learning-based AST-Encoding for Cross-platform Binary Code Similarity Detection // DSN. 2021.
6. Hou F., Jansen S. A survey of the state-of-the-art approaches for evaluating trust in software ecosystems // Journal of Software: Evolution and Process. 2024

Кубрин Г.С., Зегжда Д.П.

Санкт-Петербургский политехнический университет Петра Великого, Санкт-Петербург

ВЫЯВЛЕНИЕ КОМПЛЕКСНЫХ УЯЗВИМОСТЕЙ В ПРОГРАММНОМ ОБЕСПЕЧЕНИИ НА БАЗЕ АНАЛИЗА ГРАФОВЫХ ПРЕДСТАВЛЕНИЙ КОДА

Внедрение методов статического и динамического анализа в процессы непрерывной интеграции и поставки современного программного обеспечения приводит к уменьшению количества уязвимостей распространенных классов среди уязвимостей с высокой оценкой критичности в базах данных NIST CVE и ФСТЭК БДУ [1, 2]. Для выявляемых в последние годы критических уязвимостей ПО характерна комплексная структура, объединяющая набор единичных уязвимостей и не деструктивных воздействий на ПО в рамках целенаправленного воздействия, приводящего к значительным нарушениям свойств безопасности. Примерами критических уязвимостей, обладающих комплексной структурой, могут выступать уязвимости библиотеки ImageMagic, приводящие к нарушению свойств безопасности использующих ее веб-приложений, и уязвимость ПО cPanel, требующая комплексного воздействия на подсистему управления сеансами с целью достижения обхода проверки авторизации (CVE-2026-41940) [4, 5].

Для описания компонентов комплексных уязвимостей ПО предлагается использовать известные системы описания единичных уязвимостей различных классов на

базе CVSS-метрики критичности и дополнительную модель описания VDG (Vulnerability Description Graph) для характеристики задействованных в рамках эксплуатации ресурсов вычислительной системы, в рамках которой функционируют модули уязвимого ПО [3]. Для описания не деструктивных воздействий на модули ПО, используемых в рамках комплексной уязвимости, предлагается использовать классификацию на базе функциональной роли воздействия:

- воздействия, направленные на раскрытие некритической информации, необходимой для эксплуатации единичных уязвимостей. Например, воздействия данного типа могут быть направлены на раскрытие системного времени ОС, в рамках которой функционирует ПО, с целью упрощения перебора идентификаторов критических ресурсов, формируемых на базе временной метки;
- воздействия, направленные на внедрение подконтрольных данных в ресурсы вычислительной среды с низким уровнем целостности. Например, воздействия данного типа могут быть направлены на внедрение подконтрольных данных, необходимых для эксплуатации единичных уязвимостей на базе ошибок класса «path traversal»;
- воздействия, направленные на внедрение конфиденциальных данных в косвенно подконтрольный сегмент ресурсов вычислительной системы. Например, воздействия данного типа могут быть направлены на внедрение критических аутентификационных данных в формируемые веб-страницы, содержащие уязвимости утечки данных.

Для выявления единичных уязвимостей и модулей ПО, подверженных не деструктивным воздействиям, потенциально применимым в рамках комплексных уязвимостей, предлагается использовать ансамбль алгоритмов анализа графовых представлений кода ПО с различным уровнем детализации: граф зависимости компонентов ПО, граф зависимости процедур по операциям вызова, граф зависимости по данным [6].

В рамках работы предлагается подход к выявлению комплексных уязвимостей ПО на базе оркестрации LLM-агентов с обеспечением доступа к файлам восстановленного кода ПО, к набору инструментов статического анализа графовых представлений и к инструментам динамического анализа (рисунок 1). Предлагаемая архитектура системы анализа защищенности ПО содержит следующие компоненты:

- узел с LLM-моделью на базе ПО Ollama используется для локального развертывания LLM-моделей [7];
- LLM-агент на базе ПО OpenCode предоставляет функционал многоагентной системы с возможностью доступа к файлам исследуемого ПО и набору локальных инструментов [8];
- расширяемая локальная база знаний заполняется специфической для задач анализа защищенности информацией с возможностью доступа со стороны LLM-агента (например, описанием известных уязвимостей, характерных для исследуемого класса ПО);
- набор DAST-, SAST-инструментов, представленный штатными средствами ПО OpenCode (утилиты командной строки ОС Linux) и расширенный набором инструментов, реализующих алгоритмы анализа графовых представлений кода ПО;
- стенд с экземпляром исследуемого ПО используется для предоставления многоагентной системе экземпляра объекта исследования с целью обеспечения возможности использования DAST-инструментов и с целью выполнения итоговой верификации потенциальных уязвимостей.

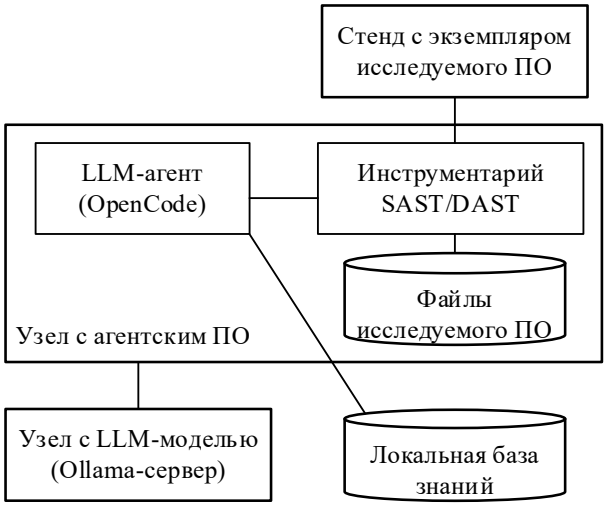


Рисунок 1 – Схема многоагентной системы выявления комплексных уязвимостей на базе инструментов Ollama и OpenCode

Предлагаемый подход учитывает известные недостатки LLM-моделей, связанные с большим количеством ложноположительных срабатываний вследствие генерации не верифицированных фактов в выходной последовательности модели, путем использования DAST-инструментов в рамках обращения к подагенту с целью генерации доказательства корректности на базе работы утилиты. Подход адаптирован для анализа ПО без доступа к исходным кодам на базе дополнительного набора инструментов, автоматизирующих декомпиляцию программных модулей (скомпилированных программных модулей, Java- и .NET-модулей). Слой оркестрации предназначен для выполнения семантического поиска потенциальных комплексных уязвимостей в множестве комбинаций единичных уязвимостей и не деструктивных воздействий за счет накопления промежуточных результатов анализа в базе знаний и учета контекста ПО.

Библиографический список:

1. База данных угроз безопасности информации // ФТЭК [Электронный ресурс]. Режим доступа: https://bdu.fstec.ru/vul (дата обращения: 23.04.2026).
2. База данных уязвимостей NVD // NIST [Электронный ресурс]. Режим доступа: https://nvd.nist.gov/vuln (дата обращения: 23.04.2026).
3. Кубрин Г.С., Зегжда Д.П. Формальная модель описания комплексных уязвимостей программного обеспечения с использованием аппарата теории графов //МЕТОДЫ И ТЕХНИЧЕСКИЕ СРЕДСТВА ОБЕСПЕЧЕНИЯ БЕЗОПАСНОСТИ ИНФОРМАЦИИ Учредители: Санкт-Петербургский политехнический университет Петра Великого. – №. 34. – С. 206-208.
4. Техническое описание уязвимостей безопасности библиотеки ImageMagic [Электронный ресурс]. Режим доступа: https://pwn.ai/blog/imagemagick-from-arbitrary-file-read-to-rce-in-every-policy-zeroday (дата обращения: 12.04.2026).
5. Уязвимость обхода авторизации ПО cPanel и WHM CVE-2026-41940 // NIST [Электронный ресурс]. Режим доступа: https://nvd.nist.gov/vuln/detail/CVE-2026-41940 (дата обращения: 03.05.2026).
6. Kubrin G. S., Zegzhda D. P. Detecting Defects in Multicomponent Software Using a Set of Universal Graph Representations of the Code //Automatic Control and Computer Sciences. – 2024. – Т. 58. – №. 8. – С. 1255-1262.
7. Репозиторий инструмента Ollama [Электронный ресурс]. Режим доступа: https://github.com/ollama/ollama (дата обращения: 11.02.2026).
8. Репозиторий инструмента OpenCode [Электронный ресурс]. Режим доступа: https://github.com/anomalyco/opencode (дата обращения: 11.02.2026).

Пагуба Г.Ю., Кондачков Е.Д.
Санкт-Петербургский политехнический университет Петра Великого, Санкт-Петербург

ПОДХОД К АВТОМАТИЗАЦИИ ПОИСКА МЕХАНИЗМОВ ВНЕШНЕГО ВЗАИМОДЕЙСТВИЯ ВСТРАИВАЕМОГО ПРОГРАММНОГО ОБЕСПЕЧЕНИЯ

С точки зрения развития информационных технологий последние годы выделяются большим ростом количества мобильных устройств, в частности мобильных устройств под управлением ОС Android [1], а также количеством личной и конфиденциальной информации, которая хранится в этих устройствах. Мобильное устройство сегодня является для человека не только средством связи, но и удобным мобильным хранилищем для персональной информации: документов, паролей, токенов и прочих артефактов, которые каким-либо образом указывают на личность человека и раскрывают о нём информацию. В следствие глубокой интеграции личных данных пользователей в компактный форм-фактор, мобильное устройство становится выгодной и важной мишенью для злоумышленника. Количество подходов к получению данных с устройства жертвы постоянно растет, но, вероятно, наиболее опасными являются техники физического подключения и использования низкоуровневых уязвимостей прошивки для получения полного доступа к системе в обход любых средств защиты [2][3][4]. На основании этого факта одним из приоритетных направлений в области обеспечения безопасности смартфонов является своевременное обнаружение уязвимостей в низкоуровневых компонентах, таких как загрузчики и ядро операционной системы, а также прошивки периферийных контроллеров, например модема устройства.

Процесс обнаружения включает в себя автоматизированное тестирование образца исследуемого ПО при различных условиях его работы и вариациях входных данных. В данной работе в качестве образца ПО рассматриваются различные виды первичного загрузчика смартфона, выполняющие проверку загружаемых в память образов и выполняющие функции восстановления прошивки устройства по некоторому внешнему интерфейсу. Таким образом для формирования корректных тестовых данных необходимо проведение анализа и формирования общей характеристики реализованного формата сообщений и функций его обработки. Информация об используемых внешних интерфейсах исследуемого образца ПО, формате передаваемых и принимаемых сообщений и функциях их обработки называется профилем внешнего взаимодействия ПО в контексте данной работы. Формирование описанного профиля внешнего взаимодействия является одним из основных этапов исследования встраиваемого ПО.

Для формирования профиля используется представление исследуемого исполняемого файла в виде теоретико-множественной модели, в которой в качестве множеств выделяются следующие: множество функций F_0, множество адресов памяти M_0, множество констант C_0.

Для формирования теоретико-множественной модели используется анализ архитектурных спецификаций [5] и фрагментов программного кода, полученных из открытых источников. На основе проведённого анализа для каждого из рассматриваемых интерфейсов формируются целевые множества конфигурационных констант C_i, множества функций F_i, множества адресов M_i и множества строк S_i.

Формирование множеств, в частности множеств F_i, сопряжено с сложностью поиска и дальнейшего корректного сопоставления с реальными образцами, поэтому для упрощения анализа целевые множества дополняются примерами из ранее проанализированных исполняемых файлов.

По окончании формирования исследуемых множеств F_0, S_0, M_0 и целевых множеств F_i, M_i, S_i производится сопоставление этих множеств и поиск пересечений для выявления искомых элементов. Результатом сопоставления является множество функций,

которые являются наиболее вероятными кандидатами на роль методов, используемых в работе рассматриваемых интерфейсов внешнего взаимодействия.

Для обеспечения возможности сопоставления различных реализаций одного и того же участка исходного кода в описываемом подходе используется хранение и сопоставление этих участков в виде промежуточного представления (intermediate representation, IR). Данный подход позволяет производит сопоставление исследуемых и целевых множеств вне зависимости от оптимизаций программного кода, применяемых компилятором и различий в поколениях архитектуры (например, ARM v8 и ARM v9).

Таким образом, в работе предложен метод автоматизированного поиска механизмов внешнего взаимодействия программного кода для встраиваемого программного обеспечения на основе теоретико-множественной модели и представления анализируемого кода в виде промежуточного представления.

Список использованных источников:

1. Android Usage Statistics 2026 [Электронный ресурс]. URL: https://www.demandsage.com/android-statistics/. – (дата обращения: 10.05.2026).
2. Сайт «Лаборатория Касперского» [Электронный ресурс]. URL: https://www.kaspersky.ru/about/press-releases/laboratoriya-kasperskogo-obnaruzhila-uyazvimost-v-chipsetah-snapdragon. – (дата обращения: 10.05.2026).
3. How To Tame Your Unicorn: Exploring And Exploiting Zero-Click Remote Interfaces of Huawei Smartphones [Электронный ресурс]. URL: https://i.blackhat.com/USA21/Wednesday-Handouts/US-21-Komaromy-How-To-Tame-Your-Unicorn-wp.pdf. – (дата обращения: 11.05.2026).
4. Android Data Encryption in depth [Электронный ресурс]. URL: https://blog.quarkslab.com/android-data-encryption-in-depth.html. – (дата обращения: 09.05.2026).
5. ARM v8-A Machine readable architecture specification [Электронный ресурс]. URL: https://developer.arm.com/-/media/developer/products/architecture/armv8-a-architecture/A64_v82A_ISA_xml_00bet3.1.tar.gz. – (дата обращения: 10.05.2026).

Прокопенко Д.Н.[1], Грибков Н.А.[2]

[1]ГУМРФ им. адмирала С.О.Макарова, Санкт-Петербург

[2]Санкт-Петербургский политехнический университет Петра Великого, Санкт-Петербург

СНИЖЕНИЕ ЛОЖНОПОЛОЖИТЕЛЬНЫХ СРАБАТЫВАНИЙ В СТАТИЧЕСКОМ АНАЛИЗЕ БЕЗОПАСНОСТИ JAVA-КОД

Статический анализ безопасности применяется для раннего выявления дефектов в жизненном цикле разработки, однако для Java-проектов его практическая ценность часто ограничивается ложноположительными срабатываниями [1]. Проблема усиливается из-за внедрения зависимостей (dependency injection) - механизма, при котором объект получает необходимые зависимости из внешнего контейнера или конфигурации, а не создает их внутри себя, а также из-за ORM-слоев, методов-оберток, проектных валидаторов и абстракций Java-фреймворков. В условиях CI/CD и DevSecOps избыточный поток предупреждений снижает доверие разработчиков к анализатору и затрудняет приоритизацию исправлений [2]. Кроме того, исследования безопасности CI/CD-конвейеров показывают актуальность автоматизированного анализа поведения сборщика и выявления потенциально вредоносной активности в процессе сборки [3].

Цель работы - разработать и апробировать подход, при котором LLM помогает извлекать семантические признаки из Java-кода, но итоговая оценка строится на проверяемой taint-цепочке: source, промежуточные преобразования, sink, санитизаторы и условия достижимости. Такой подход особенно важен для прикладных систем, где требования к надежности и безопасности ПО связаны с эксплуатацией критически значимых объектов и транспортной инфраструктуры [4].

Проект VTC реализован на Python 3.10+ и оформлен как CLI-инструмент vtc analyze. Конфигурация задается через env и параметры командной строки: пользователь выбирает LLM-провайдера OpenAI или Ollama, режим верификации, формат вывода и директорию с Java-кодом. Архитектура разделена на четыре стадии: LLM-инференс спецификации, построение графа потоков данных, верификация цепочек и генерация объяснения результата.

На первой стадии AST/regex-парсер извлекает методы, сигнатуры, импорты, строки кода и контекст класса. Затем LLM формирует структурированную JSON-спецификацию sources, sinks и sanitizers. Чтобы снизить риск галлюцинаций, результат сопоставляется с реальными идентификаторами, строками и фрагментами метода. Дополнительно применяется классификатор source/sink, учитывающий типичные Java-паттерны пользовательского ввода, файловых операций, командной строки, SQL-запросов, HTML-вывода и сетевых вызовов.

На второй стадии строится граф потоков данных: узлы соответствуют переменным, выражениям и вызовам методов, а ребра описывают присваивания, возвраты, передачу параметров и межметодные связи. Для поиска путей используются BFS и A*, при наличии внешнего инструментария возможна интеграция с Joern/Code Property Graph [5]. На третьей стадии цепочка получает статус verified, false или unverifiable. Быстрый режим проверяет достижимость по CFG, строгий режим дополнительно использует символьное выполнение и SMT-ограничения Z3. На четвертой стадии формируется объяснение с указанием source, sink, промежуточных узлов, CWE и рекомендации по исправлению.

Первичная апробация проводилась на прикладных тестовых наборах из реальных проектов: Keycloak, Spark, Jenkins Perfecto, Jenkins Docker Commons, JSPWiki, cron-utils и Spring Framework. Всего оценивались десять Java-файлов с семнадцатью ожидаемыми цепочками. В текущем снимке VTC подтвердил пять ожидаемых цепочек, включая Path Traversal в Spark и Command Injection в Jenkins Perfecto. Одновременно были зафиксированы тридцать восемь ложноположительных и одиннадцать пропущенных случаев; значительная часть результатов осталась в статусе не классифицированных.

Полученные результаты показывают, что прототип уже способен строить объяснимые taint-цепочки и отделять часть подтверждаемых потоков от шумовых предупреждений. При этом метрики показывают, что учет особенностей работы Java-фреймворков пока реализован не полностью: для Keycloak и других сложных проектов требуется расширение межфайлового анализа, правил методов-оберток, распознавания санитизаторов и калибровки LLM-промптов. Следовательно, LLM в данном подходе не заменяет формальную проверку, а используется как компонент, который помогает сформировать гипотезу о потенциальной цепочке и передать ее на верификацию [6].

По сравнению с простым LLM-поиском уязвимостей проект имеет более строгую структуру результата: каждая находка привязана к конкретному source, sink, пути распространения данных, типу CWE и механизму проверки. Это делает отчет пригодным для DevSecOps-процессов, где важны воспроизводимость, трассируемость и возможность отличить подтверждаемую цепочку от предположения модели.

Таким образом, в работе предложен гибридный подход к снижению ложноположительных срабатываний в статическом анализе Java-кода. Программный проект VTC объединяет LLM-инференс, графовый анализ, поиск путей BFS/A*, CFG-проверку, Z3-ограничения и генерацию объяснимого отчета. Апробация на прикладных тестовых наборах из реальных проектов подтвердила работоспособность пайплайна, но

также выявила необходимость дальнейшего развития межфайлового анализа, базы Java-фреймворков и правил распознавания санитизаторов. Дальнейшее развитие проекта связано с расширением набора верификационных правил, добавлением экспертной обратной связи и использованием результатов в CI/CD-контуре безопасной разработки.

Список литературы:

1. Livshits V.B., Lam M.S. Finding Security Vulnerabilities in Java Applications with Static Analysis // USENIX Security Symposium. 2005. P. 271-286.
2. Ныркov А.П., Прокопенко Д.Н. О методологии разработки приложений - «Непрерывная интеграция и доставка» // Региональная информатика (РИ-2024): материалы XIX Санкт-Петербургской международной конференции. Санкт-Петербург, 2024. С. 223-224.
3. Бугаев В. А., Жуковский Е. В., Лырчиков А. А. Обнаружение потенциально вредоносной активности в конвейерах CI/CD на основе анализа поведения сборщика // Проблемы информационной безопасности. Компьютерные системы. 2025. № 1. С. 69-82. DOI: 10.48612/jisp/at5b-46tf-zet9.
4. Прокопенко Д.Н. DevSecOps-подход к разработке ПО для автономных транспортных систем // Информационная безопасность регионов России (ИБРР-2025): материалы XIV Санкт-Петербургской межрегиональной конференции. Санкт-Петербург, 2025. С. 267-269.
5. Yamaguchi F., Golde N., Arp D., Rieck K. Modeling and Discovering Vulnerabilities with Code Property Graphs // IEEE Symposium on Security and Privacy. 2014. P. 590-604.
6. Chen M. et al. Evaluating Large Language Models Trained on Code. 2021.

УДК 004.056

Самарин Н.Н.[1], Мигачев М.В.[2]

[1] Ордена Трудового Красного Знамени федеральное государственное бюджетное образовательное учреждение высшего образования «Московский технический университет связи и информатики», Москва

[2] Федеральное государственное унитарное предприятие «Научно-исследовательский институт «Квант», Москва

ВОЗМОЖНОСТИ АВТОМАТИЗАЦИИ ПОИСКА УЯЗВИМОСТЕЙ ПРОГРАММНОГО ОБЕСПЕЧЕНИЯ

Стремительное увеличение объёма и сложности программного обеспечения (ПО) в современных информационных системах закономерно сопровождается ростом числа дефектов безопасности, потенциально пригодных для эксплуатации злоумышленниками. По данным каталога CVE, ежегодно фиксируется более двадцати тысяч новых уязвимостей, при этом значительная их часть относится к категориям высокой и критической степени опасности [1].

Традиционно под уязвимостью программного обеспечения понимается ошибка в исходном коде, бинарном представлении или архитектуре программы, эксплуатация которой приводит к нарушению свойств конфиденциальности, целостности или доступности обрабатываемой информации [2]. Автоматизированный поиск уязвимостей предполагает совокупность методов и инструментальных средств, обеспечивающих выявление подобных недостатков без необходимости непосредственного участия эксперта в анализе каждого фрагмента кода.

Выделяется три основных направления автоматизированного анализа: статический, динамический и гибридный анализ. Известные средства статического анализа применяют методы абстрактной интерпретации и символьного исполнения, что позволяет обнаруживать широкий спектр уязвимостей, включая переполнения буфера, разыменования нулевых указателей, инъекции и небезопасную работу с памятью [3-5]. Существенным недостатком статических методов является высокая доля ложных срабатываний, обусловленная особенностями применяемых абстракций.

Динамический анализ, в свою очередь, основан на наблюдении за поведением программы в процессе её исполнения с заданными входными данными. Наиболее результативной разновидностью динамического анализа в последние годы признано фаззинг-тестирование, предполагающее автоматизированную генерацию потока структурированных и неструктурированных входных данных с последующим выявлением аномальных состояний. Прикладная значимость данного подхода подтверждена результатами проекта OSS-Fuzz, посредством которого обнаружено свыше десяти тысяч уязвимостей в проектах с открытым исходным кодом.

Особое место в ряду методов автоматизированного анализа занимает символьное исполнение. В отличие от классического тестирования, символьное исполнение трактует входные данные как символьные переменные и тем самым исследует пути исполнения программы, формируя для каждого из них условие достижимости в виде формулы логики первого порядка [4]. Тем не менее, практическое применение символьного исполнения существенно ограничено проблемой комбинаторного взрыва числа путей, недостаточной поддержкой системных вызовов и операций с плавающей точкой.

Отдельного рассмотрения заслуживает применение методов машинного обучения и искусственных нейронных сетей в задаче поиска уязвимостей [5]. Современными научными работами предложен ряд архитектур, в которых исходный или бинарный код представляется в виде последовательности токенов, графа или их комбинации, после чего обученная модель производит классификацию фрагментов кода по наличию уязвимости. Экспериментальные исследования показывают, что графовые нейронные сети, обученные на представлениях, объединяющих информацию о потоке управления и потоке данных, превосходят традиционные инструменты статического анализа по метрике F-меры на ряде эталонных наборов данных, включая SARD и Devign [6].

Выявленные ограничения отдельных подходов привели к появлению гибридных методологий, объединяющих несколько техник анализа в рамках единого решения. Типичная архитектура такого средства предполагает первичную локализацию подозрительных фрагментов средствами статического анализа, последующее направленное фаззинг-тестирование этих фрагментов и подтверждение эксплуатируемости посредством символьного исполнения [7]. Сочетание методов обеспечивает снижение количества ошибок первого и второго рода при сохранении приемлемой полноты обнаружения и в настоящее время рассматривается как наиболее перспективное направление развития средств анализа защищённости ПО.

Несмотря на изученность проблемы, некоторые вопросы сохраняют актуальность. К ним относятся: ограниченная масштабируемость глубоких методов анализа на проекты промышленного масштаба, недостаточная эффективность известных средств применительно к встроенному ПО и микроконтроллерам, слабая поддержка анализа многопоточных и распределённых программ. Отдельной задачей является обеспечение воспроизводимости результатов и формирование репрезентативных эталонных наборов данных, пригодных для объективной сравнительной оценки разрабатываемых инструментов.

Дальнейшее повышение результативности автоматизированного анализа связано с разработкой гибридных методологий, интегрирующих преимущества формальных и эмпирических методов, с применением больших языковых моделей при условии их верификации формальными средствами, а также с формированием стандартизованных

эталонных наборов данных, обеспечивающих объективную сравнительную оценку существующих и перспективных решений.

Список литературы:

1. CVE Program. Common Vulnerabilities and Exposures: official catalogue [Электронный ресурс]. URL: https://www.cve.org (дата обращения: 15.05.2026).
2. Chess B., West J. Secure Programming with Static Analysis. Boston: Addison-Wesley Professional, 2020. 624 p.
3. Manès V. J. M., Han H., Han C., Cha S. K., Egele M., Schwartz E. J., Woo M. The Art, Science, and Engineering of Fuzzing: A Survey // IEEE Transactions on Software Engineering. 2021. Vol. 47, № 11. P. 2312–2331.
4. Baldoni R., Coppa E., D'Elia D. C., Demetrescu C., Finocchi I. A Survey of Symbolic Execution Techniques // ACM Computing Surveys. 2018. Vol. 51, № 3. Article 50. P. 1–39.
5. Zhou Y., Liu S., Siow J., Du X., Liu Y. Devign: Effective Vulnerability Identification by Learning Comprehensive Program Semantics via Graph Neural Networks // Advances in Neural Information Processing Systems. 2019. Vol. 32. P. 10197–10207.
6. Khare A., Dutta S., Li Z., Solko-Breslin A., Alur R., Naik M. Understanding the Effectiveness of Large Language Models in Detecting Security Vulnerabilities // arXiv preprint arXiv:2311.16169. 2023. 18 p.
7. Stephens N., Grosen J., Salls C., Dutcher A., Wang R., Corbetta J., Shoshitaishvili Y., Kruegel C., Vigna G. Driller: Augmenting Fuzzing Through Selective Symbolic Execution // Proceedings of the Network and Distributed System Security Symposium (NDSS). San Diego, 2016. P. 1–16.

Татаренко Д.Г., Бондаренко В.С., Киселёв А.Н.

Военно-космическая академия имени А.Ф. Можайского, Санкт-Петербург

ГИБРИДНЫЙ ПОДХОД К ВЫЯВЛЕНИЮ СКРЫТЫХ ВРЕДОНОСНЫХ ФУНКЦИЙ В OPEN-SOURCE ПРОЕКТАХ НА ОСНОВЕ СТАТИЧЕСКОГО АНАЛИЗА И LLM-ИНТЕРПРЕТАЦИИ

Введение. Расширение использования open-source программного обеспечения сопровождается ростом числа инцидентов, связанных с внедрением вредоносного кода в открытые репозитории. По данным компании Sonatype, количество вредоносных пакетов в open-source экосистемах превысило 700 тыс. в 2024 году, продемонстрировав рост более чем на 150% за год, а к 2025 году совокупное число таких пакетов превысило 1,2 млн [1].Традиционные средства статического анализа позволяют выявлять известные уязвимости, однако не обеспечивают интерпретацию результатов и требуют экспертной оценки. При этом значительная часть предупреждений таких инструментов относится к ложноположительным срабатываниям, что усложняет их практическое применение. В этих условиях актуальной является задача разработки систем, способных автоматически выявлять потенциально опасные конструкции и формировать понятные объяснения выявленных угроз.

Основная часть

Постановка проблемы и обоснование подхода

Существующие методы анализа программного кода обладают ограничениями: статические анализаторы ориентированы на сигнатуры, а языковые модели без дополнительного контекста склонны к недостоверным выводам.

Предлагаемый гибридный подход заключается в совместном применении трёх методов, каждый из которых компенсирует недостатки остальных:

Статический анализ обеспечивает выявление потенциально опасных конструкций в программном коде на основе формальных правил и сигнатур, позволяя обнаруживать известные уязвимости и подозрительные паттерны;

RAG-конвейер обеспечивает сопоставление обнаруженных фрагментов кода с описаниями известных уязвимостей и сценариев атак из базы знаний, формируя контекст для дальнейшего анализа;

LLM-блок формирует интерпретируемое заключение на естественном языке на основе результатов анализа и найденного контекста, обеспечивая объяснение характера выявленной угрозы и рекомендаций по её устранению.

Таким образом, подход реализует принцип комплементарности: статический анализ отвечает за обнаружение потенциально опасных конструкций в программном коде, retrieval-компонент - за сопоставление выявленных фрагментов с известными уязвимостями и сценариями атак, а генеративная модель - за синтез интерпретируемого заключения на основе ограниченного и обоснованного контекста.

Реализация подхода: архитектура программной системы

Для практической апробации предложенного подхода разработана программный комплекс анализа open-source репозиториев. Его архитектура включает три основных функциональных модуля:

4. **Модуль статического анализа:** осуществляет автоматический анализ исходного кода с использованием инструментов статического анализа (Semgrep, Bandit, Gitleaks).

На данном этапе выявляются потенциальные уязвимости, утечки чувствительных данных и подозрительные конструкции.

5. **Модуль поиска и сопоставления:** для обеспечения интерпретируемости результатов используется механизм сопоставления выявленных фрагментов кода с описаниями известных уязвимостей и сценариев атак. Входные данные (описания выявленных уязвимостей) преобразуются в векторное представление с использованием модели Sentence-Transformers, после чего осуществляется поиск схожих записей в векторной базе знаний (Chroma), содержащей описания уязвимостей и типовых сценариев атак.

6. **LLM-модуль (интерпретация результатов):** локальная языковая модель, развернутая с использованием среды Ollama, формирует итоговое заключение на основе переданного контекста. Модель генерирует интерпретируемое объяснение, включающее: описание потенциальной угрозы, обоснование её выявления, оценку уровня риска, рекомендации по устранению.

Серверная часть реализована на FastAPI, пользовательский интерфейс - на React. Хранение данных осуществляется локально с использованием векторной базы(Chroma).

Выводы. Предложенный подход позволяет повысить эффективность анализа безопасности программного кода за счёт объединения методов обнаружения и интерпретации. В отличие от традиционных средств, программный комплекс обеспечивает не только выявление потенциальных уязвимостей, но и формирование обоснованных объяснений. Реализация программного комплекса подтверждает применимость подхода для задач аудита open-source проектов и обеспечения информационной безопасности.

Список литературы:

1. Sonatype: Появление уязвимостей в цепочке поставок программного обеспечения. [Электронный ресурс]. — URL: https://www.sonatype.com/state-of-the-software-supply-chain/2026/open-source-malware.

Шестаков И.А., Бирюков Д.Н.
Военно-космическая академия имени А.Ф. Можайского, Санкт-Петербург

МЕТОД ВЫЯВЛЕНИЯ НЕБЕЗОПАСНЫХ СОСТОЯНИЙ ANDROID-ПРИЛОЖЕНИЙ НА ОСНОВЕ АНАЛИЗА КОДА И КОНФИГУРАЦИИ

Развитие экосистемы Android-приложений сопровождается увеличением количества уязвимостей и небезопасных состояний, приводящих к несанкционированному доступу к персональным данным и компрометации пользовательских устройств. Согласно OWASP Mobile Top 10, наиболее распространёнными остаются проблемы небезопасной конфигурации приложения, неправильное хранение данных и недостаточная криптографическая защита данных[6]. По данным исследования AppSec Solutions (2024), 88,6% популярных Android-приложений содержат уязвимости высокого или критического уровня Традиционные средства автоматизированного анализа, как правило, используют либо статический, либо динамический подход изолированно, без взаимной корреляции результатов, что приводит к снижению точности выявления уязвимостей [2]. В связи с этим актуальна разработка комбинированного метода, объединяющего статический (SAST) и динамический (DAST) анализ с формализованным методом выявления небезопасных состояний.

Предлагаемый метод выявления небезопасных состояний в Android-приложениях базируется на системе Mobilnik, реализующей четырёхэтапную последовательность анализа. Методологической основой служат «Методика оценки угроз безопасности информации» ФСТЕК России [5], рекомендации OWASP MASTG [7] и результаты комплексного анализа SAST-инструментов для Android [3]. Выбор подхода обусловлен необходимостью учитывать, как статические дефекты кода, видимые без запуска приложения, так и динамические уязвимости, проявляющиеся только в рабочей среде [4].

Модули системы Mobilnik. На этапе статического анализа (SAST) задействуют пять специализированных модуль: ManifestAnalyzer – проверка 7 параметров манифеста (Intent-фильтры, разрешения, Content Providers); CodeAnalyzer – 11 шаблонов для обнаружения WebView, SQL-инъекций, XSS; CryptoAnalyzer – 7 проверок криптоалгоритмов и жёстких ключей; SecretScanner – 15 типов секретов (API-ключи, токены, сертификаты); BinaryAnalyzer – 5 проверок native-компонентов. Динамический анализ (DAST) включает три модуля: StaticDASTAnalyzer (ADB, Monkey, Logcat), TrafficAnalyzer (mitmproxy, Frida, Burp Suite) и ProactiveScanner (Frida, customscripts).

Принятие решений и оценка риска. После дедупликации результатов SAST/DAST и построения многомерного вектора признаков X = (x1, ..., xn) система вычисляет вероятность небезопасного состояния P(НБС | X). Итоговая функция риска: Risk = f(attack_vector, exploit_complexity, confidence, false_positive_probability). Классификация выполняется на основе байесовского решающего правила с матрицей потерь (асимметрия ошибок), обеспечивая минимизацию среднего риска при принятии решения «Небезопасно» / «Безопасно». Привязка найденных дефектов к OWASP Mobile Top 10 (2024), MITRE ATT&CK for Mobile (v14) [6] и CWE обеспечивает соответствие отраслевым стандартам.

Рисунок 1. Структура метода анализа уязвимостей Android-приложений

Таким образом, предлагаемый метод объединяет лучшие практики статического и динамического анализа приложений, применяя байесовский подход для минимизации вероятности ложных срабатываний. В отличие от существующих решений, система обеспечивает полное покрытие уязвимостей благодаря синергии SAST- и DAST-модулей, корреляции дефектов и количественной оценки вероятности небезопасного состояния (ВКНС). В условиях постоянного роста числа и сложности мобильных угроз подобный инструмент необходим для проактивной защиты информации в Android-экосистеме.

Список литературы:

1. AppSec Solutions. Исследование уязвимостей мобильных приложений 2024. – URL: https://appsec-sting.ru/research/mobile-application-vulnerabilities (дата обращения: 28.04.2026).
2. Хромова А.Р., Петросян Л.Э. Анализ уязвимостей в системах безопасности данных // Инженерный вестник Дона. 2023. № 6. – URL: ivdon.ru/ru/magazine/archive/n6y2023/8447 (дата обращения: 15.04.2026).
3. Zhu J., Li K., Chen S. et al. A Comprehensive Study on Static Application Security Testing (SAST) Tools for Android // arXiv. 2024. – URL: https://arxiv.org/abs/2410.20740 (дата обращения: 20.04.2026).
4. Котенко И.В., Саенко И.Б., Паращук И.Б. и др. Аналитическая обработка больших массивов данных о событиях кибербезопасности // Программные продукты и системы. 2024. Т. 37. № 4. С. 487–494.
5. Методический документ «Методика оценки угроз безопасности информации». Утв. ФСТЭК России 5 февраля 2021 г. – М.: ФСТЭК, 2021. – С. 10–15, 83–85.
6. ArikkatD.R., VinodP. etal. DroidTTP: Mapping Android Applications with TTP for Cyber Threat Intelligence // arXiv. 2025. – URL: https://arxiv.org/abs/2503.15866 (датаобращения: 25.04.2026).
7. OWASP Mobile Application Security Testing Guide (MASTG). – URL: https://mas.owasp.org/MASTG (датаобращения: 13.04.2026).

5. КРИПТОГРАФИЧЕСКИЕ МЕТОДЫ ЗАЩИТЫ ИНФОРМАЦИИ

Бугаев В.А., Александрова Е.Б.
Санкт-Петербургский политехнический университет Петра Великого, Санкт-Петербург

ПОСТРОЕНИЕ СХЕМ УПОРЯДОЧЕННОЙ КОЛЛЕКТИВНОЙ ПОДПИСИ С УЧЕТОМ ПОЛИТИК ДОСТУПА

Схемы упорядоченной коллективной электронной подписи (ЭП) позволяют не только проверять целостность сообщения и подтверждать подлинность авторства, но и порядок, в котором участники схемы формировали подпись. Такие ЭП обеспечивают аутентификацию в системах электронного документооборота, протоколах безопасной маршрутизации или при многоэтапном утверждении транзакций в блокчейн-сетях [1]. Актуальность исследования квантово-устойчивых схем ЭП подтверждается принятыми в августе 2024 года стандартами FIPS 204 и FIPS 205, а также перспективой стандартизации отечественной схемы ЭП «Шиповник», основанной на алгоритме Штерна.

Цель исследования заключается в классификации подходов к созданию схем квантово-устойчивых упорядоченных коллективных подписей и в разработке общего механизма построения таких схем. Решаемая проблема состоит в том, что на данный момент не существует механизма создания схем упорядоченной коллективной ЭП на основе заданной базовой схемы, в качестве которой могут выступать принимаемые в настоящее время постквантовые стандарты. С учетом появления новых стандартов и обнаружения атак на постквантовые схемы [2] необходимо иметь возможность оперативно синтезировать схемы ЭП с требуемыми свойствами на основе заданного стандартом математического аппарата.

В ходе анализа исследований, посвященных схемам упорядоченных ЭП, были выделены три основных подхода (таблица 1). Под носителем порядка подписания в данном случае понимается сущность, позволяющая реализовать отслеживание иерархии подписантов в схеме ЭП.

Таблица 1. Подходы к синтезу упорядоченных ЭП

Подход	Носитель порядка	Примеры схем ЭП	Особенности
Использование встроенного в математический аппарат носителя порядка подписания	Операции математического аппарата: композиция изогений, делегирование базиса	[3] – порядок задается цепочкой изогенных эллиптических кривых. [4] – порядок задается делегированием базиса решетки	Отслеживание порядка реализуется благодаря свойствам математического аппарата. Требуется базовая схема ЭП, обладающая такими свойствами.

Искусственное добавление носителей порядка подписания в протокол	Коэффициенты агрегации, частичные подписи	[5, 6] – порядок задается на основе последовательно вычисленных коэффициентов агрегации. В качестве базовых протоколов взяты схема ЭП в мультипликативной группе конечного поля и протокол CSI-FiSh соответственно	Подход позволяет наделить свойством упорядоченности широкий набор базовых схем ЭП. Отсутствуют специфичные требования к математическому аппарату
Модельно-ориентированное описание порядка подписания	Внешняя модель процесса: метаграф, политика	[7] – порядок задается метаграфом и цепочкой эллиптических кривых-идентификаторов.	Возможна реализация разветвленных моделей контроля доступа, а также адаптивного контроля доступа в соответствии с политиками

Первый подход ситуативен и может применяться при наличии подходящего математического аппарата, изначально поддерживающего отслеживание порядка подписания. Наиболее универсальным является третий подход, реализующий системы адаптивного контроля доступа благодаря переходу от модели бизнес-процесса к схеме ЭП. Однако он сложен в реализации и при более тривиальных моделях контроля доступа оптимальным представляется применение второго подхода, позволяющего потенциально синтезировать свойство упорядоченности у широкого круга схем коллективной подписи, основанных на различных математических структурах.

Предлагается следующий механизм построения схем упорядоченной коллективной ЭП. На *первом* шаге выбирается базовая постквантовая коллективная схема ЭП Σ_{base} в зависимости от актуального стандарта или предметной области. На *втором* шаге фиксируются политики $P_i = (R, P, V)$, где R – роли, P – допустимые переходы между ролями, V – версия политики. *Третий* шаг заключается в выборе подхода к синтезу новой схемы на основе вариантов, представленных в табл. 1. На *четвертом* шаге задаются основные операции будущей схемы ЭП: генерация параметров и ключей, формирование и проверка подписи, доказательство корректности схемы. При необходимости описываются процедуры агрегирования ключей и подписей. *Пятый* шаг включает в себя анализ безопасности получившейся схемы ЭП. Типовые сценарии атак: вскрытие личного ключа пользователя, подделка сообщения, внедрение в цепочку подписи нарушителя, комбинирование подписи и изменение порядка подписи сообщения.

Выбор подхода на третьем шаге зависит от базовой схемы Σ_{base} и области применения будущей схемы ЭП. При выборе второго подхода потенциально могут быть использованы базовые схемы [8, 9], при выборе третьего подхода представляет интерес схема [10].

Список литературы:

1. Drake J. et al. Hash-based multi-signatures for post-quantum ethereum //Cryptology ePrint Archive. – 2025.
2. Castryck W., Decru T. An efficient key recovery attack on SIDH (preliminary version) //IACR Cryptol. ePrint Arch. – 2022. – Т. 2022. – С. 975.
3. Ярмак А. В. Схема иерархической аутентификации на основе изогений эллиптических кривых: дипломная работа: 10.05.01. – 2018.
4. Федичев А. В. Упорядоченная подпись на основе делегирования базиса решетки: дипломная работа: 10.05.01. – 2022.
5. Александрова Е.Б., Бугаев В.А., Терёшкина Е.А. Схема упорядоченной коллективной подписи на основе CSI-FiSh // Неделя науки ИКНК, 2024. – С. 50–53.
6. Boldyreva A. et al. Ordered multisignatures and identity-based sequential aggregate signatures, with applications to secure routing //Proceedings of the 14th ACM conference on Computer and communications security. – 2007. – С. 276-285

7. Ярмак А. В. Групповая аутентификация и криптографический контроль доступа в системах с иерархической структурой на основе изогений эллиптических кривых: диссертация на соискание ученой степени кандидата технических наук / ФГАОУ ВО «Санкт-Петербургский политехнический университет Петра Великого», 2023.
8. D'Alconzo G. et al. A Framework for Group Action-Based Multi-signatures and Applications to LESS, MEDS, and ALTEQ //IACR International Conference on Public-Key Cryptography, 2025. – С. 99-133.
9. Paul A. et al. An Efficient Sequential Aggregate Signature Scheme with Lazy Verification // Cryptology ePrint Archive. – 2025.
10. Brodsky M. F. et al. Monotone-policy aggregate signatures //Annual International Conference on the Theory and Applications of Cryptographic Techniques, 2024. – С. 168-195.

Голованов А.А., Кирьянова А.П.
Университет ИТМО, Санкт-Петербург

ИСПОЛЬЗОВАНИЕ МЕХАНИЗМА FAIL-STOP ПОДПИСИ ПРИ ВЗАИМОДЕЙСТВИИ УМНЫХ УСТРОЙСТВ В IOT-СЕТЯХ

Работа выполнена в рамках государственного задания (проект FSER-2025-0003)

Одной из важных задач при построении сетей умных устройств является управление составом группы, включающее добавление новых устройств, исключение скомпрометированных или старых устройств, а также обновление ключевого материала после изменения состава группы. Особенно существенной является задача удаления устройства из сети, поскольку после исключения оно не должно иметь возможности участвовать в последующих протоколах аутентификации и формирования подписей [1].

Для обеспечения проверяемости таких действий все критические операции внутри группы должны фиксироваться в журнале событий [2]. К таким операциям относятся аутентификация, добавление и удаление устройств, отзыв ключей, ротация группового ключевого материала и т.д. Ведение журнала событий необходимо для любой IT-системы, оно обеспечивает целостность истории событий, защиту от незаметных и несанкционированных изменений, удалений или переупорядочиваний записей, а также даёт возможность разрешения спорных ситуаций, связанных с подлинностью административных действий.

В данной работе предлагается рассмотреть новые групповые подписи на основе подписи со свойством доказательства факта подделки (fail-stop signatures, FSS) [3], как вспомогательный механизм для повышения защищённости журнала событий в IoT-сетях. В отличие от обычных цифровых подписей, FSS позволяют легитимному подписанту в случае появления валидной, но не созданной им подписи, предъявить доказательство её подделки.

В докладе продемонстрировано как использование новой схемы групповых подписей на основе FSS в задачах ведения журнала событий в IoT-сетях позволяет сочетать эффективное использование ресурсов с достаточным уровнем безопасности.

Список литературы:

1. Mangar, Ravindra, et al. "A Trigger for the Autonomous Decommissioning of Smart Devices." Proceedings of the 15th International Conference on the Internet of Things. 2025.

2. Wang Q. et al. Fear and logging in the internet of things //Network and Distributed Systems Symposium. – 2018.
3. Pedersen T. P., Pfitzmann B. Fail-stop signatures //SIAM Journal on Computing. – 1997. – Т. 26. – №. 2. – С. 291-330.

Даниленко А.Ю., Акимова Г.П.
Федеральный исследовательский центр «Информатика и управление» РАН, Москва

СКРЫТЫЕ КАНАЛЫ В СИСТЕМАХ ЭЛЕКТРОННОГО ДОКУМЕНТООБОРОТА

Одним из способов вмешательства в нормальное функционирование распределенных многопользовательских автоматизированных информационных систем (АИС) является использование скрытых каналов (СК). Данный метод отличается от других, традиционных методов тем, что такие атаки нельзя обнаружить и предотвратить с помощью «традиционных» средств зашиты информации (СЗИ) таких, как идентификация и аутентификация пользователей, управление доступом, межсетевые экраны, системы обнаружения вторжений. Это обусловлено тем, что «традиционные» СЗИ контролируют только информационные потоки, которые проходят по каналам, предназначенным для их передачи. Опасность СК для информационных технологий (ИТ), АИС и защищаемых активов организации связана с отсутствием контроля средствами защиты скрытых информационных потоков, что может привести к утечке информации, нарушить целостность информационных ресурсов и программного обеспечения в компьютерных системах или создать иные препятствия по реализации ИТ.

Системы электронного документооборота (СЭД) относятся к категории распределенных многопользовательских АИС с различными правами доступа к информационным объектам и действиям в рамках АИС, поэтому для них задача перекрытия СК утечки данных и деструктивного влияния на АИС весьма актуальна.

Согласно [1] СК подразделяются на следующие категории:

- СК по памяти;
- СК по времени;
- скрытые статистические СК.

СК по памяти основаны на наличии памяти, в которую передающий субъект записывает информацию, а принимающий - считывает ее.

СК по времени предполагают, что передающий информацию субъект модулирует с помощью передаваемой информации некоторый изменяющийся во времени процесс, а субъект, принимающий информацию, в состоянии демодулировать передаваемый сигнал, наблюдая несущий информацию процесс во времени.

Скрытый статистический канал использует для передачи информации изменение параметров распределений вероятностей любых характеристик системы, которые могут рассматриваться как случайные и описываться вероятностно-статистическими моделями.

Характерными особенностями СЭД являются наличие большого числа информационных объектов с разными правами доступа к ним, интенсивное использование каналов связи для обмена данными, в том числе за пределами контролируемых зон. Соответственно, основными направлениями атак на СЭД с использованием СК будут:

- несанкционированный доступ (НСД) к данным, в том числе доступ легально работающего пользователя (внутреннего нарушителя) к закрытой для него информации;

– передача закрытой информации внутри АИС между легально работающими пользователями, что можно рассматривать как канал связи между нарушителями, который невозможно отследить службе безопасности;
– передача закрытой информации за пределы АИС от легально работающего пользователя внешнему нарушителю, т.е. неконтролируемый канал связи между нарушителями;
– нарушение функционирования АИС.

Нарушители при реализации атак могут использовать следующие возможности: работа в качестве легального пользователя и управление программно-аппаратными закладками либо троянскими программами.

В случае СЭД возможны следующие способы реализации перечисленных атак:
– поисковый запрос на поиск документов, для которых запросивший пользователь (нарушитель) не имеет права просмотра. СЭД может быть спроектирована и реализована таким образом, что пользователю показывается количество документов, удовлетворяющих поисковому запросу, и, возможно, их список, в котором приводится значимая информация про эти документы (например, регистрационный номер, дата регистрации, краткое содержание). В этом случае нарушитель может получить большой объем информации, к которой не допущен. Тем самым, в данном случае налицо грубая ошибка проектирования АИС;
– модификация загрузки канала связи. В данном случае нарушитель может управлять загрузкой канала таким образом, чтобы передавать информацию получателю на внешнем сервере. Этот способ работоспособен даже в случае наличия криптографической защиты канала;
– модификация загрузки процессора. Управляя потоком задач, нарушитель может передавать информацию на внешний сервер, который использует программно-аппаратные средства удаленного мониторинга состояния технических средств;
– использование служебных полей в текстовых документах офисных форматов, например Word, для записи в них информации, предназначенной для внешнего нарушителя;
– перехват модифицированных файлов за пределами КЗ, т.е. атака «человек посередине». Происходит отправка текста или рисунка ничего не подозревающему партнеру, а злоумышленник перехватывает файл, расшифровывает информацию и отправляет файл адресату;
– управление троянскими программами или программно-аппаратными закладками. Таким образом можно вносить разнообразные помехи в работу АИС для ее дискредитации или полного прекращения функционирования. Более изощренная атака – это изменение списков доступа к информационным объектам, для этого троян должен запускаться с правами администратора.

Рассмотрим способы противодействия атакам на СЭД с применением СК. К таким средствам противодействия могут быть отнесены:
– проектирование и реализация АИС с учетом возможного возникновения СК в готовом программном продукте;
– тематические исследования программного кода с целью исключения программных закладок;
– специальные проверки и исследования технических средств с целью выявления аппаратных закладок и наличия побочных излучений;
– шифрование трафика, особенно за пределами контролируемой зоны;
– применение шифраторов, выравнивающих загрузку канала связи путем добавления служебных пакетов данных;
– протоколирование и регулярный просмотр протоколов безопасности;
– контроль целостности передаваемых пакетов данных.

Таким образом, основные признаки наличия СК в СЭД:

- аномальное поведение системы, в том числе снижение быстродействия;
- необъяснимое изменение загрузки процессора;
- появление в открытом доступе защищаемой информации.

Литература

1. ГОСТ Р 53113.1-2008. Информационная технология. Защита информационных технологий и автоматизированных систем от угроз информационной безопасности, реализуемых с использованием скрытых каналов. Части 1 и 2.

Корнеева В.А., Александрова Е.Б.

Санкт-Петербургский политехнический университет Петра Великого, Санкт-Петербург

ПОДПИСЬ С АДАПТАЦИЕЙ НА ОСНОВЕ ГОСТ 34.10-2018 С УЧАСТИЕМ ДОВЕРЕННОГО ПОСРЕДНИКА

В работе [1] авторами была предложена схема подписи с адаптацией на основе ГОСТ 34.10-2018, реализующая протокол атомарного двустороннего обмена без посредника (Схема 1). Подпись с адаптацией связывает публикацию подписи с раскрытием секрета. Известные конструкции [2] опираются на зарубежные алгоритмы. В данной работе предлагается Схема 2, которая дополнительно устраняет уязвимость широкополосного скрытого канала.

В стандарте ГОСТ 34.10-2018 одноразовый параметр k выбирается подписантом самостоятельно. Как показано в работе [3], это создает возможность широкополосного скрытого канала, когда подписант может закодировать в k дополнительную информацию объемом до $log_2 q \approx 256 - 512$ бит на одну подпись. В Схеме 1 данная уязвимость сохраняется, поскольку k по-прежнему выбирается одной стороной. Для ее устранения в Схему 1 вводится доверенный посредник W, участвующий в совместной выработке значения k.

Протокол совместной выработки параметра k. Перед формированием неполной подписи подписант и посредник W выполняют следующий обмен:

1. Подписант выбирает случайное $r_1 \in {Z_q}^*$, вычисляет $R_1 = r_1 P$, $K = r_1 \cdot Q_W$, где $Q_W = a_W P$ – долговременный открытый ключ посредника (известен всем участникам), и передает R_1 посреднику W.

2. Посредник W вычисляет общий секрет $K = a_W \cdot R_1$, где a_W – долговременный закрытый ключ посредника, $r_2 \equiv H(K)(mod\ q)$ и передает r_2 подписанту.

3. Подписант устанавливает $k \equiv r_1 \cdot r_2 (mod\ q)$. При $k = 0$ протокол повторяется с новым r_1.

После выработки k алгоритмы *pSign, pVrfy, Adapt* и *Ext* выполняются идентично предложенной Схеме 1 с единственным изменением, заключающемся в том, что в *pSign* шаг выбора случайного k заменяется использованием совместно выработанного значения. Формат публикуемой подписи не изменяется и совпадает с форматом ГОСТ 34.10-2018.

Устранение широкополосного скрытого канала. В модели случайного оракула при честно действующем посреднике W и вычислительной сложности задачи дискретного логарифмирования на эллиптической кривой для любого PPT-подписанта S выполняется неравенство $Pr[k = k^*] \leq \frac{q_H}{q}$, где q_H – общее число запросов S к оракулу H. При $q = 2^{256}$ и $q_H = 2^{80}$ получаем вероятность 2^{-176}.

Подписант фиксирует $R_1 = r_1 P$ до получения r_2, поэтому скорректировать r_1 под уже известное значение r_2 невозможно. Точка $K = a_W \cdot r_1 P$ не может быть вычислена

подписантом без знания a_W. Чтобы выполнялось сравнение $r_1 \cdot r_2 \equiv k^*(mod\ q)$, необходимо знать значение $r_2 \equiv k^* \cdot {r_1}^{-1}(mod\ q) \in Z_q$, совпадение с которым имеет вероятность $\frac{1}{q}$.

Границы применимости. Защита от широкополосного скрытого канала обеспечивается при условии честного поведения W. В случае сговора подписанта и W посредник раскрывает a_W, что позволяет подобрать r_1 с заранее известным значением $r_2 \equiv H(a_W \cdot r_1 P)(mod\ q)$.

Таким образом, предложенная модификация подписи с адаптацией снижает пропускную способность широкополосного скрытого канала, при этом формат публикуемой подписи остается неизменным.

Список литературы:

1. Корнеева В. А., Александрова Е. Б. Подпись с адаптацией на основе ГОСТ 34.10-2018 / Неделя Науки ИКНК. СПб: Политех-Пресс – 2026. – С.119-120.
2. Erwig A., Fuchsbauer G., Kiltz E., Loss J. Two-party adaptor signatures from identification schemes / IACR international conference on public-key cryptography. – 2021. – P. 451 – 480.
3. Клевцов А. А., Фионов А. Н. Предложение о модификации российского стандарта электронной цифровой подписи с целью ликвидации широкополосного скрытого канала / Интерэкспо Гео-Сибирь. – 2022. – Т. 6. – С. 96-100.

Маткова Е.И., Калинин М.О.

Санкт-Петербургский политехнический университет Петра Великого, Санкт-Петербург

БЕЗОПАСНОСТЬ КОНСОРЦИУМНЫХ БЛОКЧЕЙН-СИСТЕМ В РАСПРЕДЕЛЕННЫХ КИБЕРФИЗИЧЕСКИХ СРЕДАХ УМНОГО ГОРОДА

Исследование выполнено за счет гранта Российского научного фонда №24-11-20005, https://rscf.ru/project/24-11-20005/; грант Санкт-Петербургского научного фонда (договор №24-11-20005 о предоставлении регионального гранта)

В настоящее время технология блокчейн широко обсуждается как инструмент обеспечения целостности и защищенности критических данных в распределенных киберфизических системах умного города [1]. В консорциумных блокчейнах, контролируемых группой участников, добавление транзакций выполняется участниками, а защита от внедрения злоумышленником поддельных блоков реализуется через алгоритм консенсуса, основанный на голосовании доверенных узлов, что гарантирует неизменность данных и хронологии цепочки блоков. Однако наиболее популярные механизмы консенсуса PoA и PBFT консорциумных блокчейнов уязвимы к ряду атак на консенсус [2-4] (табл. 1), что осложняет внедрение блокченйа в критические информационные инфраструктуры умных городов и требует разработки более надежного подхода к защите блокчейна.

Таблица 1. Сопоставление возможности осуществления атак на консорциумные блокчейны в распределенной киберсреде умного города

Алгоритм консенсуса	Атака 51%	Эгоистичный майнинг	Eclipse	Атака Сивиллы	Двойная трата	Атака на генератор случайных чисел
PoA	–	–	+	+/–	–	+/–
PBFT	–	–	+	+	–	+/–
Разработанное решение	–	–	–	–	–	–

Для повышения защищенности и масштабируемости консорциумных блокчейнов построен новый гибридный алгоритм консенсуса (рис. 1), в котором заменен механизм стейкинга на майнинг новых блоков, внедрено создание блока не только доверенными узлами сети с сохранением права валидации только за доверенными участниками, введены ограничения на добавление новых участников, а также использована модель машинного обучения с подкреплением для формирования очередности создания блоков.

Для экспериментальной оценки разработанного решения создан макет блокчейн-системы, реализованы функции создания/валидации блоков, синхронизации цепочек, управления связями и типовые сценарии атак, направленных на алгоритмы консенсуса PoA, PBFT и разработанный алгоритм (согласно табл. 1). Например, в ходе анализа устойчивости консорциумного блокчейна при атаке Eclipse разработанный алгоритм консенсуса продемонстрировал, что новые узлы не добавляются, поскольку согласно алгоритму, необходимо добавить хотя бы один доверенный узел-сосед для нового узла. Когда узел в соседях имеет доверенный узел, он сможет получать информацию о корректной цепочке блоков через него, а не через узел злоумышленника.

В ходе анализа устойчивости к атаке Сивиллы разработанный алгоритм консенсуса продемонстрировал, что добавление нового недоверенного узла осуществимо, но в рамках работы алгоритма неэффективно, т.к. валидация происходит только доверенными участниками. Добавление доверенного узла злоумышленником невозможно, т.к. доверенные участники юридически проверяются.

При атаке на генератор случайных чисел разработанный алгоритм консенсуса продемонстрировал устойчивость за счет использования модели машинного обучения с подкреплением для формирования очередности создания блоков, которая каждые N раундов формирует новую очередь на основании активности узлов блокчен-системы.

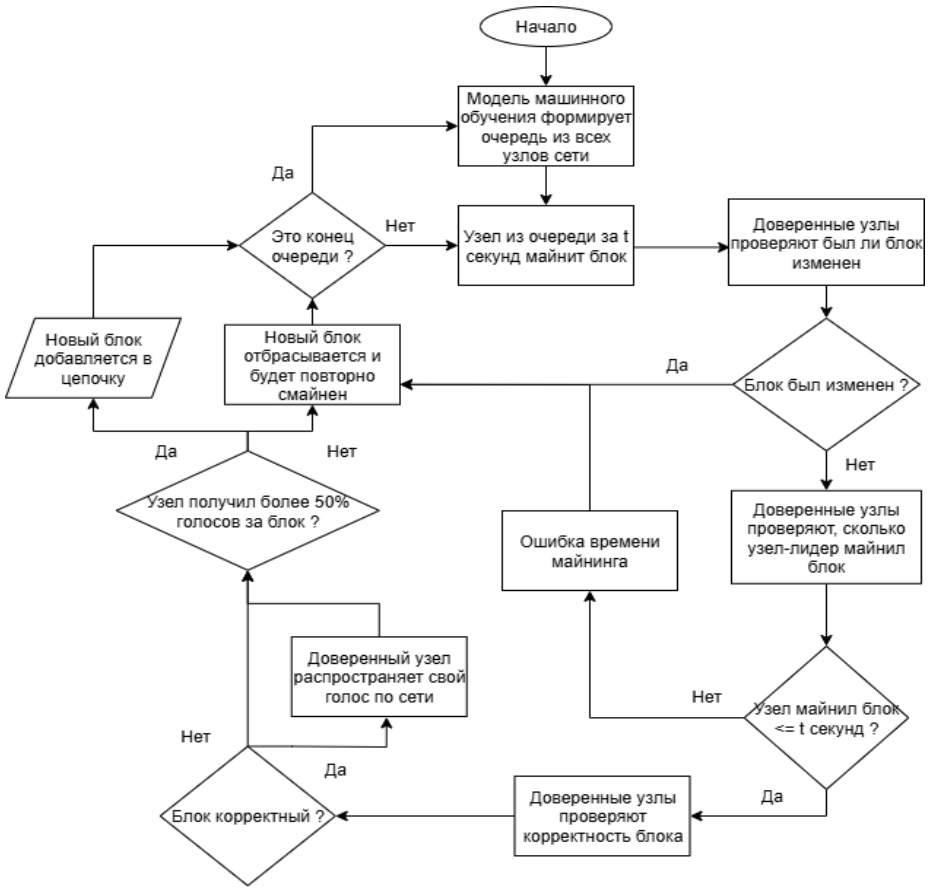


Рис. 1. Схема разработанного гибридного алгоритма консенсуса

Предложенный алгоритм обеспечивает высокий уровень устойчивости ко всему спектру атак на алгоритмы консенсуса и позволяет преодолеть основные недостатки существующих консорциумных блокчейн-решений, используемых в распределенных киберфизических средах умных городов.

Список литературы:

1. Заяц А.М. Блокчейн-системы и технология / СПб.: Лань, 2025. – 112 с.
2. Hussein, Z. Evolution of blockchain consensus algorithms: a review on the latest milestones of blockchain consensus algorithms / Z. Hussein, M.A. Salama, S.A. El-Rahman // Cybersecurity. – 2023. – Vol. 6, no. 30. – DOI: 10.1186/s42400-023-00163-y.
3. Mastromauro, D. Survey of Attacks and Defenses on Consensus Algorithms for Secure Data Replication in Distributed Systems / D. Mastromauro, M. Souto Andrade, O. Ozmen, M.A. Kinsy // IEEE Access. – 2025. - Vol. 13. – P. 143631-143667. – DOI: 10.1109/ACCESS.2025.3597260.
4. Guru, A.; Mohanta, B.K.; Mohapatra, H.; Al-Turjman, F.; Altrjman, C.; Yadav, A. A Survey on Consensus Protocols and Attacks on Blockchain Technology. Appl. Sci. 2023, 13, 2604. https://doi.org/ 10.3390/app13042604 – (дата обращения 12.11.2025)

Охлопков Н.М.
Санкт-Петербургский политехнический университет Петра Великого, Санкт-Петербург

ИНТЕЛЛЕКТУАЛЬНОЕ ФОРМИРОВАНИЕ ПРОВЕРЯЕМЫХ КОНТРОЛЬНЫХ БЛОКОВ В БЛОКЧЕЙН-СИСТЕМАХ IOT-ИНФРАСТРУКТУРЫ

Исследование выполнено за счет гранта Российского научного фонда №24-11-20005, https://rscf.ru/project/24-11-20005/; грант Санкт-Петербургского научного фонда (договор №24-11-20005 о предоставлении регионального гранта)

Блокчейн – одноранговая система фиксации транзакций без доверенного центра, в которой каждый новый блок криптографически связан с предыдущим [1]. Блокчейн широко применяется в IoT-инфраструктурах, системах промышленной автоматизации, транспортных сетях и системах умного города, где требуется обеспечить целостность критически важных данных при отсутствии единого посредника [2]. Однако использование блокчейна в IoT-среде выявляет противоречие между классической моделью полного узла и ограниченными ресурсами периферийных устройств. Узлы должны хранить и проверять растущую историю блоков, тогда как IoT-устройства имеют ограниченную память и короткие периоды сетевой активности.

Существующие подходы к снижению затрат на хранение и синхронизацию блокчейна включают локальное удаление старых блоков, облегченные клиенты, проверяемые снимки состояния и компактные криптографические доказательства [3-5]. Эти методы уменьшают отдельные виды нагрузки, но не всегда позволяют одновременно ограничивать фактический «хвост» истории, обеспечивать предсказуемое время синхронизации и формировать проверяемую точку отсчета. Структуры на основе дерева Меркла позволяют компактно фиксировать состояние реестра [6], а проверяемые снимки ускоряют загрузку узлов [7], однако момент создания такой точки обычно определяется статически. Для IoT-инфраструктуры умного города требуется адаптивный механизм, учитывающий неравномерную нагрузку.

Одним из перспективных решений является архитектура блокчейна с плавающим генезис-блоком, в которой актуальная точка отсчета периодически смещается вперед, а старый исторический префикс исключается из активного хранения [8, 9]. Такой подход сокращает объем хранимых данных и ускоряет подключение узлов. Однако ранние реализации плавающего генезис-блока в основном используют фиксированные или эвристические правила формирования контрольной точки. При всплесках транзакционной активности это приводит к несвоевременному созданию контрольного блока и нарушению ограничений по синхронизации.

Целью работы является создание нового метода предиктивного формирования проверяемых контрольных блоков в блокчейн-системе IoT-инфраструктуры умного города. Метод направлен на ограничение объема хранимой истории, повышение предсказуемости синхронизации узлов и снижение риска принятия неподтвержденной или поддельной контрольной точки. Основная идея метода состоит в том, что решение о создании нового контрольного блока принимается не только по фактическому достижению порогов, но и на основе прогноза роста «хвоста» истории. Для этого собираются признаки состояния реестра и текущей нагрузки, включая размер блока, число транзакций, объем хвоста после последнего контрольного блока, время с момента его формирования, интенсивность поступления транзакций, время синхронизации и параметры доступности валидаторов. На основе этих признаков вычислительная модель ИИ на базе линейной онлайн-регрессии оценивает будущий прирост «хвоста» истории. Онлайн-обновление параметров позволяет адаптировать модель к изменяющейся нагрузке без полного переобучения.

Ограничение синхронизации задается через максимально допустимое время, за которое новый или долго отсутствовавший узел должен получить актуальное состояние сети. Это ограничение преобразуется в допустимый объем данных с учетом средней пропускной способности канала. Если прогноз показывает риск превышения допустимого

хвоста, система заранее инициирует формирование контрольного блока. Для устойчивости к ошибкам прогноза вводится коэффициент уверенности модели. При недостаточной уверенности решение принимается по резервным условиям: числу блоков, времени с момента последней контрольной точки или фактическому объему хвоста.

Контрольный блок рассматривается не только как средство сокращения истории, но и как проверяемый объект доверия. Он включает корневой хеш состояния реестра, составной сертификат цифровых подписей валидаторов, хеш гиперпараметров модели, коэффициент уверенности прогноза и хеш предыдущего контрольного блока. Принимающий узел проверяет связность с предыдущей контрольной точкой, кворум подписей, согласованность параметров модели и достаточность коэффициента уверенности.

Экспериментальная оценка проводилась на блокчейн-цепочке со стохастически изменяющейся нагрузкой. Сравнивались классическая схема без контрольных блоков, политики по числу блоков, времени, объему данных, гибридная пороговая политика и разработанный предиктивный метод. Результаты показали, что классическая схема приводит к монотонному росту хранимых данных, а фиксированные политики остаются чувствительными к всплескам нагрузки. Гибридная политика снижает риск нарушений, но не устраняет его полностью. Разработанный метод формирует контрольные блоки чаще, однако обеспечивает более низкую верхнюю границу хранимого объема и удерживает «хвост» истории в допустимых пределах.

Таким образом, разработанный метод обеспечивает переход от статического формирования контрольных точек к адаптивному управлению «хвостом» истории блокчейна. Сочетание плавающего генезис-блока, линейной онлайн-регрессии и проверяемости контрольного блока позволяет сократить объем хранения, повысить предсказуемость синхронизации и снизить риск принятия некорректной контрольной точки в IoT-инфраструктуре умного города.

Список литературы:

1. Nakamoto S. Bitcoin: A peer-to-peer electronic cash system. 2008.
2. Uddin M. A. et al. A survey on the adoption of blockchain in iot: Challenges and solutions //Blockchain: Research and Applications. – 2021. – Т. 2. – №. 2. – С. 100006.
3. Wang Y., Wang H., Cao Y. Comprehensive review of storage optimization techniques in blockchain systems //Applied Sciences. – 2024. – Т. 15. – №. 1. – С. 243.
4. Chatzigiannis P., Baldimtsi F., Chalkias K. Sok: Blockchain light clients //International Conference on Financial Cryptography and Data Security. – Cham : Springer International Publishing, 2022. – С. 615-641.
5. Bünz B. et al. Flyclient: Super-light clients for cryptocurrencies //2020 IEEE Symposium on Security and Privacy (SP). – IEEE, 2020. – С. 928-946.
6. Wood G. et al. Ethereum: A secure decentralised generalised transaction ledger //Ethereum project yellow paper. – 2014. – Т. 151. – №. 2014. – С. 1-32.
7. Ren L., Chen W. T., Ward P. A. S. Snapshotsave: fast and low storage demand blockchain bootstrapping //Proceedings of the 36th Annual ACM Symposium on Applied Computing. – 2021. – С. 291-300.
8. Busygin A., Kalinin M., Konoplev A. Supporting connectivity of VANET/MANET network nodes and elastic software-configurable security services using blockchain with floating genesis block //SHS Web of Conferences. – EDP Sciences, 2018. – Т. 44. – С. 00020.
9. Busygin A. et al. Floating genesis block enhancement for blockchain based routing between connected vehicles and software-defined VANET security services //Proceedings of the 11th International Conference on Security of Information and Networks. – 2018. – С. 1-2.

Охлопков Н.М., Калинин М.О.
Санкт-Петербургский политехнический университет Петра Великого, Санкт-Петербург

ЗАЩИТА БЛОКЧЕЙНА УМНОГО ГОРОДА ОТ УГРОЗ БЕЗОПАСНОСТИ, ЭКСПЛУАТИРУЮЩИХ УЯЗВИМОСТИ СЧЕТЧИКОВ ВРЕМЕНИ

Исследование выполнено за счет гранта Российского научного фонда №24-11-20005, https://rscf.ru/project/24-11-20005/; грант Санкт-Петербургского научного фонда (договор №24-11-20005 о предоставлении регионального гранта)

Современные блокчейн-системы, применяемые в инфраструктуре умного города, используют метки времени для журналирования событий, упорядочивания транзакций, проверки актуальности сообщений и анализа состояния распределенной сети [1]. В IoT-среде умного города нарушение достоверности системного времени даже одного узла может привести к ошибкам локальной обработки, нарушению последовательности событий и снижению доверия к распределенному реестру. Поэтому защита подсистемы времени в блокчейн-сети становится самостоятельной задачей обеспечения безопасности.

Угрозы безопасности, связанные с нарушением достоверности времени, можно разделить на три основных класса: угрозы сетевой синхронизации, угрозы локальному времени узла и угрозы распределенного уровня [2]. На сетевом уровне основной целью атак являются протоколы синхронизации времени, прежде всего NTP. Возможны подмена ответов сервера времени, replay-атаки, искусственное внесение задержки, компрометация сервера времени и отказ внешних временных сервисов [2, 3]. На локальном уровне нарушитель может воздействовать на RTC, изменять системное время средствами операционной системы, выполнять rollback-атаки или использовать дрейф локальных часов [4, 5]. На распределенном уровне ошибка отдельного узла перестает быть локальной, если узел начинает распространять ложную временную оценку среди соседей. В результате возникает риск распространения недостоверного времени и формирования ложного доверия к некорректной временной шкале [6].

Известные методы и подходы закрывают только часть вышеуказанных угроз безопасности. NTP обеспечивает простую и дешевую синхронизацию, но не гарантирует криптографическую подлинность источника времени [2]. NTS повышает защищенность за счет аутентификации сервера и защиты целостности временных пакетов [7], однако не устраняет полностью проблему асимметричных задержек. Многосерверные методы, включая Khronos, повышают устойчивость к компрометации отдельных серверов за счет выбора и фильтрации временных источников [8]. PTP обеспечивает высокую точность, но ориентирован преимущественно на контролируемые локальные сети [9]. TPM и механизмы удаленной аттестации позволяют контролировать состояние платформы и монотонность локальных счетчиков, но не являются самостоятельным источником внешнего времени [10]. Таким образом, для IoT-ориентированного блокчейна систем умного города требуется комбинированный метод защиты, объединяющий сетевую синхронизацию, локальный аппаратный контроль и распределенную проверку времени.

Авторами разработан метод формирования доверенного системного времени узла блокчейна. Метод включает три взаимосвязанных контура: сетевой контур NTP/NTS, локальный аппаратный контур RTC+TPM и распределенный механизм взаимоконтроля узлов. Итоговая временная оценка формируется не из одного источника, а как агрегированное значение нескольких частных оценок: сетевой, RTC-оценки, монотонной TPM-оценки и распределенной оценки соседних узлов. Для каждого контура задается динамический коэффициент доверия. Если источник становится недоступным или демонстрирует аномальное поведение, его вес уменьшается вплоть до исключения из итоговой оценки.

Сетевой контур использует несколько внешних серверов времени и отбраковывает ответы, не прошедшие NTS-проверку. Далее применяется медианная фильтрация и

усреднение допустимых ответов. Локальный контур сопоставляет RTC с монотонным счетчиком TPM и выявляет признаки подмены системного времени или аномального дрейфа. Распределенный контур основан на обмене подписанными временными оценками между соседними узлами. Узлы, систематически отклоняющиеся от общей временной картины, помечаются как подозрительные и временно исключаются из взаимоконтроля.

Экспериментальная оценка выполнена на стенде, моделирующем работу узла распределенного реестра в следующих сценариях: в штатном режиме, при подмене NTP-ответов, replay-атаке, атаке искусственного внесения задержки и изоляции узла от внешней сети. Проведено сравнение четырех методов: NTP с NTS, Khronos, локальный RTC с TPM и предлагаемый комбинированный метод. Сетевые методы уязвимы к атакам задержки и потере внешней синхронизации. Локальный метод постепенно накапливает ошибку из-за дрейфа RTC. Построенное новое решение удерживает ошибку времени на меньшем уровне за счет перекрестной проверки сетевого, аппаратного и распределенного контуров.

Разработанный метод повышает устойчивость блокчейна умного города к киберугрозам, эксплуатирующим уязвимости счетчиков времени. Его применение позволяет снизить риск принятия транзакций и событий с недостоверными временными метками, обнаруживать локальные и сетевые аномалии времени и поддерживать доверенную временную шкалу даже при частичной изоляции узла.

Список литературы:

1. Khan I. et al. Blockchain applications for Internet of Things – A survey //Internet of Things. – 2024. – Т. 27. – С. 101254.
2. Anwar F. M., Srivastava M. Applications and challenges in securing time //12th USENIX Workshop on Cyber Security Experimentation and Test (CSET 19). – 2019.
3. Mizrahi T. RFC 7384: Security requirements of time protocols in packet switched networks. – 2014.
4. Liu J. et al. {TimeTravel}: Real-time timing drift attack on system time using acoustic waves //34th USENIX Security Symposium (USENIX Security 25). – 2025. – С. 3885-3902.
5. Alder F. et al. About time: On the challenges of temporal guarantees in untrusted environments //Proceedings of the 6th Workshop on System Software for Trusted Execution. – 2023. – С. 27-33.
6. Badertscher C. et al. Ouroboros chronos: Permissionless clock synchronization via proof-of-stake //Cryptology ePrint Archive. – 2019.
7. Dowling B., Stebila D., Zaverucha G. Authenticated network time synchronization //25th USENIX security symposium (USENIX security 16). – 2016. – С. 823-840.
8. Rozen-Schiff N. et al. RFC 9523: A Secure Selection and Filtering Mechanism for the Network Time Protocol with Khronos. – 2024.
9. Eidson J. C., Fischer M., White J. IEEE-1588™ Standard for a precision clock synchronization protocol for networked measurement and control systems //Proceedings of the 34th annual precise time and time interval systems and applications meeting. – 2002. – С. 243-254.
10. Birkholz H. et al. RFC 9334: Remote ATtestation procedureS (RATS) Architecture. – 2023.

Охлопков Н.М., Калинин М.О., Коноплев А.С.,
Пахомов М.А., Пахомова О.А., Ярков С.Д.
Санкт-Петербургский политехнический университет Петра Великого, Санкт-Петербург

АНОНИМИЗАЦИЯ РАСПРЕДЕЛЕННОГО РЕЕСТРА В КРИТИЧЕСКИ ВАЖНЫХ ИНФОРМАЦИОННЫХ СИСТЕМАХ НА ОСНОВЕ ОДНОРАЗОВЫХ КРИПТОГРАФИЧЕСКИХ КЛЮЧЕЙ

Исследование выполнено за счет гранта Российского научного фонда №24-11-20005, https://rscf.ru/project/24-11-20005/; грант Санкт-Петербургского научного фонда (договор №24-11-20005 о предоставлении регионального гранта)

В критически важных информационных системах распределенные реестры применяются для фиксации событий, контроля целостности данных и обеспечения проверяемости операций без централизованного посредника. Однако открытость блокчейна создает угрозу деанонимизации участников. В традиционных псевдоанонимных блокчейн-системах конфиденциальность субъектов находится под постоянной угрозой из-за развития специальных аналитических инструментов. На сегодняшний день компании, занимающиеся блокчейн-аналитикой (Chainalysis, Elliptic), уже применяют методы эвристического анализа, машинного обучения и кластеризации адресов для построения графов транзакционной активности. Особенно значимыми являются эвристики общего входа, адреса сдачи, анализ временных профилей и структурная связность транзакционного графа [1, 2]. Например, в контексте инфраструктуры умного города анализ публичного графа транзакций позволяет злоумышленникам отслеживать маршруты перемещения дронов- доставщиков, выявлять паттерны энергопотребления в домах, определяя периоды отсутствия жильцов, и раскрывать коммерческую тайну производственных цепочек.

Существующие методы повышения анонимности, включая CoinJoin/CoinShuffle, zk-SNARK-подходы, CryptoNote/Monero и stealth-адреса, обеспечивают разные уровни приватности, но имеют ограничения: необходимость координации, высокая вычислительная нагрузка, увеличение размера транзакций или сложность интеграции в существующие инфраструктурные реестры [3-6]. Для критически важных систем дополнительно требуется не только анонимность, но и возможность контролируемого подтверждения владения отдельным выходом при аудите.

Представлен метод анонимизации блокчейна на основе одноразовых криптографических ключей. Вместо публикации повторяемого адреса получателя в каждой транзакции формируется новый одноразовый открытый ключ выхода. Основу составляет концепция строгого разделения ончейн- и оффчейн-процессов. В публичном пространстве (на уровне блокчейна) транзакционная активность должна выглядеть как случайный шум. Передача ценности осуществляется исключительно между одноразовыми адресами, которые генерируются для каждой новой операции и уничтожаются сразу после ее завершения. Внешний наблюдатель фиксирует лишь изолированные переводы, криптографическая и математическая связь между которыми вычислительно невыводима. Однако для обеспечения подотчетности перед легитимными контрагентами в систему внедряется механизм делегируемых метаданных и локального хранения криптографических доказательств. В момент инициации транзакции клиентское устройство локально формирует вспомогательный набор данных (цифровую квитанцию), который криптографически связывает одноразовый адрес отправителя с идентификатором платежа. Эти вспомогательные данные не транслируются в публичный реестр, а сохраняются в защищенной памяти на стороне клиента. При возникновении необходимости в аудите (например, при сверке расчетов с поставщиком коммунальных услуг), пользователь предоставляет эти доказательства уполномоченной стороне по закрытому оффчейн-каналу. Проверяющая сторона, получив метаданные, получает возможность математически валидировать принадлежность конкретного публичного перевода данному пользователю.

Таким образом, формируется архитектура нулевого разглашения для неограниченного круга лиц, которая при этом полностью удовлетворяет требованиям легитимного аудита, сохраняя способность узлов отчитываться о совершенных действиях перед авторизованными участниками сети.

Вся архитектурная нагрузка по обеспечению приватности перенесена непосредственно на сторону клиента, внутрь самого периферийного устройства. Защита должна формироваться до того, как пакет данных вообще попадет в публичную сеть. В рамках такого подхода клиентское программное обеспечение берет на себя всю работу: оно локально генерирует новые одноразовые ключи, автономно дробит исходящие суммы и формирует серию параллельных микропереводов. Для внешнего наблюдателя и систем мониторинга блокчейна эти действия выглядят как самые обычные базовые переводы между ничем не примечательными пользователями. Не происходит никаких вызовов сложной ончейн-логики, алгоритмическое запутывание следов реализуется исключительно за счет клиентских решений на устройстве. Это позволяет сохранить сетевые комиссии на уровне копеечных транзакций и полностью избежать маркировки активов как подозрительных. В условиях жестко ограниченных бюджетов и вычислительных мощностей оборудования умного города автономная клиентская фрагментация остается технически и экономически оправданным решением.

Метод дополняется регламентом жизненного цикла одноразовых ключей, селективным доказательством владения конкретным выходом для доверенного аудитора и политикой UTXO shaping. Последняя уменьшает долю мультивходовых транзакций и ослабляет эвристику общего входа. Эксперимент на имитационном UTXO-реестре показал снижение F1-меры успешной деанонимизации до 0,171. Таким образом, предложенный метод снижает связываемость транзакций и может использоваться для повышения приватности распределенных реестров в критически важных информационных системах.

Список литературы:

1. Meiklejohn S. et al. A fistful of bitcoins: characterizing payments among men with no names //Proceedings of the 2013 conference on Internet measurement conference. – 2013. – С. 127-140.
2. Reid F., Harrigan M. An analysis of anonymity in the bitcoin system //Security and privacy in social networks. – New York, NY: Springer New York, 2012. – С. 197-223.
3. Stütz R. Et al. Adoption and actual privacy of decentralized coinjoin implementations in bitcoin //Proceedings of the 4th ACM Conference on Advances in Financial Technologies. – 2022. – С. 254-267.
4. Sasson E. B. et al. Zerocash: Decentralized anonymous payments from bitcoin //2014 IEEE symposium on security and privacy. – IEEE, 2014. – С. 459-474.
5. Van Saberhagen N. CryptoNote v 2.0. – 2013.
6. Herskind L., Katsikouli P., Dragoni N. Privacy and cryptocurrencies—A systematic literature review //IEEE Access. – 2020. – Т. 8. – С. 54044-54059.

Петренко А.С.
Военно-космическая академия имени А.Ф. Можайского, Санкт-Петербург

МЕТОД ПОСТРОЕНИЯ ПОСТКВАНТОВЫХ АЛГЕБРАИЧЕСКИХ АЛГОРИТМОВ ЭЦП С ДВУМЯ СКРЫТЫМИ ГРУППАМИ

Исследование выполнено за счет гранта Российского научного фонда № 24-41-04006, https://rscf.ru/project/24-41-04006/

Введение. Появление релевантного или практически значимого квантового компьютера (ожидается к 2030 году) приведет к взлому большей части современных криптографических алгоритмов, стойкость которых базируется на вычислительной трудности решения задач факторизации и дискретного логарифмирования по простому модулю. В первую очередь уязвимыми окажутся системы шифрования и протоколы с открытым ключом, в том числе цифровые финансовые активы: механизмы и подсистемы в блокчейн-системах, предоставляющие права на цифровые объекты или данные, а также более сложные, например, смарт-контракты и цифровые деньги, в частности, национальная платформа цифрового рубля.

В настоящее время в криптографии основные усилия разработчиков криптоалгоритмов сосредоточены на разработке и стандартизации постквантовых криптографических механизмов, которые останутся актуальными даже после появления практически значимых квантовых компьютеров. Так, Национальный институт стандартов и технологий США (National Institute of Standards and Technology, NIST) по результатам конкурса среди 69 алгоритмов-кандидатов по выбору лучших квантово-устойчивых криптографических схем инкапсуляции ключа и цифровой подписи (2017-2024 гг.) подготовил первые четыре стандарта.

В России также проводятся подобные исследования и разработки. Так, экспертами-криптографами российской компании «Криптонит» разработан алгоритм электронной подписи «Шиповник», устойчивый к атакам с использованием квантового компьютера. Модификации алгоритма показывают различные результаты в скорости формирования и проверки ключей и подписи при различной криптостойкости, а разработка и сертификация алгоритма «Шиповник» ещё не завершены.

При этом угрозы, создаваемые квантовыми компьютерами для современной криптографии, всё чаще приобретают практический характер. Злоумышленники уже осуществляют сбор зашифрованных данных критической важности для последующей расшифровки после появления криптоаналитически релевантных квантовых компьютеров. В тоже время системы, использующие криптографические алгоритмы в своём составе ежегодно предъявляют всё большие требования к ресурсоёмкости, а количество атак на упомянутые системы и отдельные криптопримитивы неуклонно растёт, что в совокупности с вышеперечисленным определяет важность и значимость разработки новых постквантовых криптопримитивов.

Таким образом, актуальность темы настоящей работы объясняется появлением новой квантовой угрозы безопасности информации, ростом требований безопасности к критической информационной инфраструктуре РФ, ростом структуры и сложности поведения упомянутых систем, и уже недостаточностью (неспособностью) известных технологий (моделей, методов и средств) обеспечения информационной безопасности и киберустойчивости для решения задач обнаружения, нейтрализации и упреждения квантовых атак злоумышленников.

Возможное решение. В настоящей работе была решена научно-техническая задача понижения ресурсоёмкости постквантовых ЭЦП путём сокращения длины подписи и ключей при заданной криптостойкости для обеспечения применимости в составе блокчейн-экосистем и платформ и повышения квантовой устойчивости упомянутых систем [1-3].

Предложенный автором метод построения постквантовых ЭЦП позволил снизить ресурсоёмкость постквантовых ЭЦП путём сокращения длины подписи более чем в 6 раз и ключей не менее чем на 55% по сравнению с ближайшими аналогами при заданной криптостойкости для обеспечения применимости в составе блокчейн-экосистем и платформ, благодаря чему была повышена квантовая устойчивость упомянутых систем в 3 раза и более.

В ходе работы были получены следующие результаты:

Метод построения постквантовых алгебраических алгоритмов ЭЦП с двумя скрытыми группами на основе КНАА. В частности, обоснован выбор конечных некоммутативных ассоциативных алгебр (КНАА) с глобальной единицей в качестве математической основы, сформулирована модель с двумя скрытыми группами и единым проверочным уравнением, исключающая прямые редукции к коммутативным случаям, задан алгебраический носитель, доказана корректность разработанного алгоритма и показано «покрытие» порядка ≈p^3 значений одноразовых экспонент при ускоренном умножении;

Постквантовые алгебраические алгоритмы ЭЦП на основе вычислительной трудности решения больших систем степенных уравнений в виде класса из трёх алгоритмов ЭЦП с двумя скрытыми группами. В частности, предложен приём выделения скалярного множителя, позволяющий вдвое уменьшить разрядность показателей степени при сохранении стойкости. Также разработан способ оценивания стойкости постквантовой ЭЦП на основе КНАА с двумя скрытыми группами к атакам на основе известных подписей с использованием декомпозиции КНАА на конечные коммутативные кольца, пересекающиеся строго в множестве скалярных векторов;

Методика построения стойких постквантовых ЭЦП с малыми размерами подписи и открытого ключа на основе КНАА с двумя скрытыми группами. В частности, сформированы наборы параметров со стойкостью не ниже 2^200 и 2^256 операций соответственно категориям, аргументирована устойчивость к атакам Шора и Гровера, определены профили применения и двухэтапная стратегия миграции для существующих блокчейн-систем;

Архитектура программного комплекса на основе разработанного метода построения постквантовых ЭЦП на основе КНАА с двумя скрытыми группами для обеспечения квантовой устойчивости национальных блокчейн-экосистем и платформ.

Заключение. Полученные в ходе работы научно-технические результаты могут использоваться для повышения защищённости и устойчивости отечественных блокчейн-экосистем и платформ от квантовых угроз информационной безопасности. Разработанный метод построения постквантовых алгоритмов ЭЦП, устойчивых к атакам с использованием квантовых компьютеров, разработанные алгоритмы, методика и архитектура их применения, направленные на сокращение длин подписи и ключей и повышения устойчивости блокчейн-платформ и экосистем, могут быть применены при выполнении научных исследований в области постквантовой криптографии, а также при разработке новых защищённых блокчейн-систем и иных высоконагруженных полносвязных систем.

В дальнейшем предлагается разработать ряд методов построения гибридных криптографических схем, сочетающих постквантовые алгоритмы ЭЦП с классическими криптосистемами для обеспечения совместимости, модификацией существующих и созданием новых QUBO-оптимизаторов, проведением более детального квантового криптоанализа для существующих блокчейн-систем, а также окончательную сертификацию разработанного семейства ЭЦП.

Список литературы:

1. Алексей Петренко, «Построение постквантовых алгебраических алгоритмов ЭЦП с двумя скрытыми группами» (научная монография). – Санкт-Петербург: Изд. Питер, 2026. - 252 с. (РИНЦ).

2. Алексей Петренко, «Квантово-устойчивый блокчейн» (научная монография). – Санкт-Петербург: Изд. Питер, 2023. - 384 с. (РИНЦ), https://www.piter.com/collection/biznes-literatura/product/kvantovo-ustoychivyy-blokcheyn
3. Алексей Петренко, «Квантовая угроза технологии блокчейн». Учебно-методическое пособие. – Санкт-Петербург: Изд. Дом «Афина», 2022. - 105 с. (РИНЦ)

Прокофьева А.Э., Александрова Е.Б.
Санкт-Петербургский политехнический университет Петра Великого, Санкт-Петербург

ОКОННАЯ АГРЕГАЦИЯ ЖУРНАЛА СОСТОЯНИЙ ПРОВЕРЯЕМОЙ БАЗЫ ДАННЫХ

Рост объёмов данных повышает требования к проверяемости и целостности баз данных. В 2025 г. объём создаваемых данных в мире достигает порядка 181 зеттабайт [1], что усиливает актуальность механизмов компактного аудита и проверки истории изменений данных.

В работе [2] рассматривается архитектура постквантовой проверяемой базы данных (БД), основанной на ортогональной интеграции схемы цифровой подписи на решётках с механизмом восстановления по подсказкам в модель проверяемой базы данных Qedb [3] с целью аутентификации состояния данных. Протокол обеспечивает постквантовую аутентификацию состояния данных и защиту от replay-атак и атак отката состояния БД. Однако при практическом применении такой архитектуры возникает проблема хранения и верификации журнала состояний базы данных. В системе каждое обновление состояния сопровождается формированием новой эпохи и новой цифровой подписи для сообщения вида $\leftarrow \mathrm{enc}(\mathrm{epochi} \parallel \mathrm{Di})$, где: epochi – идентификатор эпохи (номер или версия состояния БД); Di – digest состояния БД в данной эпохе; $\mathrm{enc}(\cdot)$ – детерминированная функция кодирования данных перед подписью. В результате растёт число подписей и метаданных доверия. Это приводит к увеличению затрат памяти, времени аудита и вычислительных расходов на проверку истории изменений базы данных, что особенно критично для систем с частыми обновлениями и требованиями к аудиту и прослеживаемости изменений. В таких условиях хранение отдельной подписи для каждой эпохи становится плохо масштабируемым и требует разработки механизма компактного представления и проверки криптографически связанной цепочки состояний базы данных.

Для уменьшения накладных затрат предлагается агрегировать не сами подписи, а журнал эпох. Такой подход обусловлен тем, что стандартная BLS-подобная агрегация для схемы [4] неприменима, поскольку функция восстановления по подсказкам не является гомоморфной. Поэтому предлагаемое решение – оконная агрегация состояний БД.

Пусть окно эпох задаётся интервалом $[a, b]$. Для него формируется вектор записей журнала – $\mathrm{r}_{a,b} = (\mathrm{r}_a, \mathrm{r}_{a+1}, \ldots, \mathrm{r}_b)$. Над данным вектором строится аутентифицированная структура данных – $\rho_{a,b} \leftarrow \mathrm{VC.Commit}(\mathrm{r}_{a,b})$ – короткое открытое векторное обязательство ко всему вектору записей журнала, где $\mathrm{VC.Commit}(\cdot)$ – функция, фиксирующая всю последовательность переходов внутри окна $[a, b]$, но не требующая включать все записи

$r_a, \ldots, r_b$ в подписываемое сообщение и раскрывать элементы напрямую. Для отдельной записи r_i, где $i \in [a, b]$, формируется криптографическое свидетельство принадлежности

Для проверки агрегированного журнала сначала проверяется подпись контрольной точки

Для формальной проверки предложенного протокола использовалось средство анализа криптографических протоколов Scyther [5]. В модели рассматривались роли владельца данных, проверяющего и недоверенной БД сервера-хранилища. Проверялось, может ли

Таким образом, предложенная схема заменяет набор подписей $(\sigma_a, \sigma_{a+1}, \ldots, \sigma_b)$ одной подписью $\Sigma_{a,b}$ на корень аутентифицированной структуры $\rho_{a,b}$, что позволяет сократить число хранимых и проверяемых подписей с $b - a + 1$ до одной подписи на окно эпох и тем самым уменьшить объём метаданных доверия, необходимых для аудита журнала состояний

при этом механизм восстановления с подсказками не нарушается, поскольку функция владелец данных формирует сообщение контрольной точки окна $\mu_{a,b} \leftarrow enc(a \| b \| \rho_{a,b} \|$

случайное одноразовое значение, генерируемое проверяющим перед началом проверки, и вычисляет одну подпись - $\Sigma_{a,b} \leftarrow Sign(sk, u_{a,b})$. Итоговая контрольная точка окна имеет вид: $\sigma_{a,b} \leftarrow (a, b, \rho_{a,b}, \Sigma_{a,b})$.

вызывается только для одной стандартной подписи Σa,b, а не для агрегата нескольких подписей. Поэтому корректность подписи [4] переносится на подпись контрольной точки, а

Список литературы:

1. ig Data Statistics: How Much Data Is There in the World? [Электронный ресурс]. – 2025. – HYPERLINK "https://rivery.io/blog/big-data-statistics-how-much-data-is-there-in-the-
2. Александрова Е.Б., Прокофьева А.Э. Интеграция постквантовой подписи в проверяемую базу данных для обеспечения доверия к состоянию данных / Неделя науки ИКНК: материалы Всероссийской научно-практической конференции. 2026. С. 115-117.
3. Qedb: Expressive and Modular Verifiable Databases (without SNARKs) / Cryptology ePrint Archive. – 2025. – Report No. 2025/1408.
4. Александрова Е.Б., Прокофьева А.Э. Интеграция восстановления с подсказками в схему подписи на основе задачи обучения с ошибками / Проблемы информационной безопасности. Компьютерные системы. 2025. № S2. С. 10-22.

5. Cremers C. Scyther Tool: Verification, Falsification, and Analysis of Security Protocols [Электронный ресурс]. – URL: https://people.cispa.io/cas.cremers/downloads/scyther/scyther-manual-v1.3.0.pdf.

Самарева Д.М., Александрова Е.Б.
Санкт-Петербургский политехнический университет Петра Великого, Санкт-Петербург

ПОСТРОЕНИЕ ДОКАЗАТЕЛЬСТВА ПРИНАДЛЕЖНОСТИ ИНТЕРВАЛУ НА БАЗЕ РЕШЕТОЧНОЙ СХЕМЫ ОБЯЗАТЕЛЬСТВА С ОКРУГЛЕНИЕМ

Одной из актуальных задач криптографии является возможность доказывать свойства скрытых значений без раскрытия самих данных. Подобные доказательства могут использоваться в различных сценариях: при проведении конфиденциальных транзакций необходимо доказывать, что скрытая сумма не является отрицательной; в электронном голосовании необходимо доказывать, что голос имеет допустимое значение. Для этого используются доказательства с нулевым разглашением, в частности доказательства принадлежности интервалу [1].

Доказательства принадлежности интервалу позволяют доказывающему убедить проверяющего, что скрытое значение m лежит в некотором интервале $[l, u]$ без раскрытия самого m. Общий подход к построению подобных доказательств следующий: чтобы доказать, что значение m лежит между нижней границей l и верхней границей u, достаточно показать корректность неравенств $m - l \geq 0$ и $u - m \geq 0$. При этом сами значения $m - l$ и $u - m$ не передаются проверяющему в открытом виде, так как они могут раскрыть информацию о самом m. В доказательствах промежутка зачастую используется битовое разложение [2], позволяющее впоследствии доказывать корректность арифметических выражений.

В данной работе исследуется построение доказательства принадлежности интервалу для разработанной в [3] схемы обязательства с округлением. Данная схема построена в кольце $Z_q[X]/(X^d + 1)$, при этом особенностью схемы является использование детерминированного округления вместо случайного шума.

Б

ы

л

и

Доказывающий рассматривает две разности: $m - l$ и $u - m$. Если значение m действительно лежит в интервале, то обе эти разности являются неотрицательными. Каждая из разностей представляется в двоичном беззнаковом виде (число N выбирается таким образом, чтобы его было достаточно для представления любого значения из рассматриваемого диапазона). Доказывающий должен показать существование таких битов x_i и y_i, что:

$$m - l = \sum_{i=0}^{N-1} 2^i x_i, \quad u - m = \sum_{i=0}^{N-1} 2^i y_i,$$

$x_i, y_i \in \{0,1\}$. Тогда сумма $\sum_{i=0}^{N-1} 2^i x_i$ принимает [4] разработанная схема основ $2^N - 1$. Следовательно, доказывающий должен показать, что разность $m - l$ необходимо показать, что

Для этого все значения представляются в виде обязательств. Обозначим создание обязательства с помощью преобразования $Com()$. Таким образом, исходное сообщение m зафиксировано в обязательстве $Com(m, r)$, где r – секретный вектор. Для каждого бита x_i доказывающий строит обязательство $Com(x_i, r_i)$, где r_i – секретный вектор, используемый для сокрытия каждого бита x_i. Аналогично для разности $u - m$ строится дополнительное обязательство $Com(y_i, s_i)$, где s_i – другой секретный вектор, используемый для сокрытия битов. Значение l, отвечающее за границу интервала, является открытым, поэтому записывается как $Com(l, 0)$. Правая часть разности, равная 0, при создании обязательства выглядит как обязательство к 0 с использованием линейной комбинации секретных векторов $Com(0; r - \sum 2^i r_i)$. Таким образом, для доказательства принадлежности необходимо доказать равенство

$$Com(m; r) - Com(l; 0) - \sum_{i=0}^{N-1} 2^i \, Com(x_i; r_i) = Com(0; r - \sum_{i=0}^{N-1} 2^i \, r_i).$$

Аналогично, для верхней границы необходимо доказать равенство

$$Com(m; r) - Com(l; 0) - \sum_{i=0}^{N-1} 2^i \, Com(y_i; s_i) = Com(0; -r - \sum_{i=0}^{N-1} 2^i \, s_i).$$

Построенное доказательство может быть формально интегрировано в процедуру электронного голосования как проверка допустимости скрытого голоса. Например, для голосования с двумя вариантами можно использовать интервал $[0,1]$, где каждое значение соответствует

Таким образом, доказательство принадлежности интервалу сводится к проверке корректности двух линейных отношений над обязательствами: одно из них подтверждает корректное представление разности $m - l$, второе – разности $u - m$. При этом проверяющий не получает открытых значений самого m и убеждается только в корректности связей между обязательствами. За счет этого построенное доказательство может использоваться как основа для дальнейшего построения постквантовых протоколов конфиденциальной проверки данных.

Список литературы:

1. Christ M. et al. SoK: Zero-knowledge range proofs // Cryptology ePrint Archive. – 2024.
2. Bünz B. et al. Short proofs for confidential transactions and more // IEEE S&P, Bulletproofs. – 2018.
3. Александрова Е.Б., Самарева Д.М. Схема обязательства на решетках для доказательства линейных отношений между скрытыми значениями // Проблемы информационной безопасности. Компьютерные системы. 2025. № S2. С. 23–33. DOI: 10.48612/jisp/f7p4-n9p1-gtt6.
4. Александрова Е.Б., Самарева Д.М. Корректность доказательства линейных отношений в решеточной схеме обязательства с контролируемой погрешностью округления // Неделя науки ИКНК: материалы Всероссийской научно-практической конференции. 2026. С. 123–124.

Шенец Н.Н.

Санкт-Петербургский политехнический университет Петра Великого, Санкт-Петербург

ОБ ОДНОПРОХОДНЫХ ПРОТОКОЛАХ АУТЕНТИФИКАЦИИ С ТЕОРЕТИКО-ИНФОРМЦИОННОЙ СТОЙКОСТЬЮ

Большинство известных протоколов односторонней аутентификации относятся к типу «запрос-ответ» (challenge-response) и являются лишь вычислительно стойкими. Это значит, что для их взлома (поиска секретов участников или выдачи себя за одного из участников) требуется решить некоторую трудную математическую задачу. Задача считается трудной, если не существует полиномиального алгоритма ее решения. Совсем недавно появились работы [1 – 3], в которых авторы предлагают протоколы с теоретико-информационной стойкостью, т.е. их взлом теоретически невозможен даже для нарушителя с неограниченными вычислительными ресурсами. Однако мы выяснили, что доказательства стойкости не являются верными, и протоколы не обладают заявленным свойством безопасности. Однако сам по себе метод аутентификации из работы [1] привлекателен по следующим причинам:

- протокол однопроходный, т.е. клиент просто посылает одно сообщение серверу;
- не требуется наличие третьей доверенной стороны;
- низкие накладные расходы на вычисления;
- использование схем разделения секрета, что редко встречается в протоколах такого типа.

В данной статье мы обобщаем метод аутентификации из [1], и строго анализируем возможность обеспечения теоретико-информационной стойкости в указанных условиях. В [1] авторы предложили сразу 2 протокола аутентификации. В обоих случаях клиент и сервер разделяют между собой общий ключ *key* и текущее значение счетчика *ctr*, которое синхронно увеличивается с каждым сеансом. Однопроходность обеспечивается за счет того, что клиент доказывает серверу знание общего секрета $s = key + ctr$ (также рассматривается вариант $s = = key \| ctr$) путем передачи ему двух одинаковых, но зашифрованных по-разному значений. Основное отличие второго протокола от первого заключается в том, что в нем клиент и сервер владеют не самим значением *key*, а частичными секретами key_1 и key_2, соответственно, полученными путем применения (2, 2)-пороговой схемы разделения секрета Шамира [4]. Поэтому мы рассмотрим только первый из них.

Протокол аутентификации из [1]

Клиент		**Сервер**
		Проверяет, что $d_i / b_i \neq 0, d_i / \varepsilon_{i_1} \neq 0;$
$\varepsilon_i, b_i, d_i \xleftarrow{R} \mathbf{Z}_p^*;$		$\frac{d_i}{\varepsilon_i} = \frac{d_i}{\varepsilon_{i_1}} \times \frac{1}{\varepsilon_{i_0}};$
$\varepsilon_i = \varepsilon_{i_0} \cdot \varepsilon_{i_1} = Recover\left(\varepsilon_i^0, \varepsilon_i^1\right);$		$d_i A_0 = b_i \left(key + ctr\right) \cdot \frac{d_i}{b_i} - \left(key + ctr\right) \cdot \frac{d_i}{\varepsilon_i} \cdot \varepsilon_i^0;$
$M = \varepsilon_{i_0} \,\|\| \, \varepsilon_i^0 \,\|\| \, \varepsilon_i^1 \left(key + ctr\right) \,\|\|$	$\xrightarrow{M}$	
$\|\| \, b_i \left(key + ctr\right) \,\|\| \, d_i / b_i \,\|\| \, d_i / \varepsilon_{i_1}.$		$d_i A_1 = b_i \left(key + ctr\right) \cdot \frac{d_i}{b_i} - \left(key + ctr\right) \cdot \frac{d_i}{\varepsilon_i} \cdot \varepsilon_i^1;$
		Проверяет, что $Recover(d_i A_0, d_i A_1) = 0.$

Здесь через *Recover*() обозначена функция восстановления секрета по схеме Шамира; ε_i — случайное число, выбираемое клиентом для подтверждения знания общего с сервером секрета $s = key + ctr$; $\left(\varepsilon_i^0, \varepsilon_i^1\right)$ — пара частичных секретов, полученных путем применения схемы Шамира к ε_i; $\varepsilon_{i_0}, \varepsilon_{i_1} \in \mathbf{Z}_p^*$ — два случайных числа, произведение которых равно ε_i. Через '/' обозначена операция деления (умножения на обратный элемент в Z_p), '||' – символ конкатенации (склейки) слов.

При детальном рассмотрении данного протокола можно выделить следующие его ключевые особенности:

1. В качестве «шифрования» используется умножение открытого текса (числа) на секрет s в $\mathbf{Z}_p$. Это, фактически, мультипликативный совершенный шифр Вернама.
2. При шифровании в качестве открытого текста используется частичный секрет от (2, 2)-пороговой схемы разделения секрета. При этом сервер вместе с зашифрованными сообщениями получает также и по одному из частичных секретов для каждого из сообщений, которые не использовались при шифровании. Использование разделения секрета необходимо для того, чтобы не шифровать одно и то же, пусть и случайное, число на одном и том же ключе – ведь в этом случае невозможно получить разные шифртексты.

Заметим, что, во-первых, вместо проверки равенства двух значений в общем случае сервер может проверять равенство вида $a_2 = f(a_1)$ для двух переданных ему в зашифрованном виде значений a_1 и a_2 и произвольной известной функции f. Во-вторых, в конечном поле $\mathbf{Z}_p$. вместо одного можно применить два разных шифра Вернама — аддитивный и мультипликативный. Наконец, можно использовать не один и тот же ключ s, а два разных ключа, известных обоим участникам. В итоге, мы получаем следующую обобщенную схему аутентификации:

Обобщенная схема аутентификации

Клиент		**Сервер**
$\varepsilon \xleftarrow{R} \mathbf{Z}_p^*; \delta = f(\varepsilon);$		1. *Находит* ε_2 *и* $\delta_2;$
$\left(\varepsilon_1, \varepsilon_2\right) \leftarrow Share_1\left(\varepsilon\right); \left(\delta_1, \delta_2\right) \leftarrow Share_2\left(\delta\right)$	$\xrightarrow{M}$	2. $\varepsilon = Recover_1\left(\varepsilon_1, \varepsilon_2\right), \delta = Recover_2\left(\delta_1, \delta_2\right.$
$M = \varepsilon_1 \,\|\| \, \delta_1 \,\|\, \varepsilon_2 \cdot key_1 \,\|\| \, \delta_2 + key_2.$		3. *Проверяет, что* $\delta = f(\varepsilon).$

В докладе мы представим результаты анализа стойкости предлагаемой обобщенной схемы однопроходной аутентификации на основе разделения секрета. В частности, мы рассматриваем случай, когда ключи key_1 и key_2 зависимы (т.е. когда один получается преобразованием из другого, либо оба получается преобразованиями исходного общего

секрета s). Оказывается, что тогда достичь теоретико-информационной стойкости невозможно. Поэтому необходимо, чтобы ключи key_1 и key_2 были независимыми случайными величинами. Однако оказалось, что этого условия недостаточно для теоретико-информационной стойкости обобщенной схемы аутентификации.

Список использованных источников:

1. K. Iwamura and A. A. A. M. Kamal, "Secure User Authentication With Information Theoretic Security Using Secret Sharing-Based Secure Computation," in *IEEE Access*, vol. 13, pp. 9015-9031, 2025, doi: 10.1109/ACCESS.2025.3526632
2. A. A. A. Mohd Kamal and K. Iwamura, “Conditionally secure multiparty computation using secret sharing scheme for n <2k-1 (Short Paper),” in *Proc. 15th Annu. Conf. Privacy, Secur. Trust (PST)*, Aug. 2017, pp. 225–2255.
3. K. Iwamura and A. A. A. M. Kamal, “Communication-efficient secure computation of encrypted inputs using (k, n) threshold secret sharing,” IEEE Access, vol. 11, pp. 51166–51184, 2023.
4. A. Shamir, “How to share a secret,” Commun. ACM, vol. 22, no. 11, pp. 612–613, Nov. 1979.

Шенец Н.Н.

Санкт-Петербургский политехнический университет Петра Великого, Санкт-Петербург

АНАЛИЗ «СОВЕРШЕННЫХ» ПРОТОКОЛОВ АУТЕНТИФИКАЦИИ TUS

В 2025 году вышла статья представителей Токийского научного университета [1], в которой авторы впервые предложили методы аутентификации клиента перед сервером по открытому каналу связи, обладающие теоретико-информационной (абсолютной) стойкостью. Помимо этого, к преимуществам предложенных ими протоколов стоит отнести однопроходность (выполняется всего один раунд пересылок от клиента к серверу), отсутствие необходимости в третьей доверенной стороне и низкую вычислительную сложность. Свои методы (протоколы) авторы назвали TUS-методы. Название произошло от аббревиатуры университета (Tokyo University of Science). Стоит сказать, что изначально команда ученых под руководством профессора Кейчи Ивамуры разработала методы и протоколы многосторонних безопасных вычислений [2 – 3], которые и стали основой их методов из [1].

При детальном рассмотрении предложенных TUS-методов из [1] можно понять, что в их основе лежат следующие идеи:

3. Использовать зашумление общего секрета (пароля) с помощью сложения по модулю 2 со случайным числом либо умножения в Z_p на случайное ненулевое число. Такое зашумление равносильно совершенному шифрованию Вернама.
4. Переслать серверу в «зашифрованном» совершенным шифром виде дважды одно и то же число, чтобы сервер мог сравнить их разность с нулем. При этом каждое из чисел «шифруется» по-разному.
5. Клиент и сервер должны быть синхронизированы и помимо пароля *key* разделять общее значение счетчика *ctr*. Для каждого очередного сеанса аутентификации необходимо использовать общий секрет $s = key + ctr$ или $s = key||ctr$, и по завершении сеанса увеличивать счетчик на 1.
6. При «шифровании» в качестве открытого текста используется частичный секрет от (2, 2)-пороговой схемы разделения секрета, а в качестве ключа – как раз значение s. Для первого сообщения используется схема разделения секрета Шамира [4], а для второго – мультипликативная схема полного согласия, причем

обе схемы применяются для одного и того же случайного числа ε. При этом сервер вместе с «зашифрованными» сообщениями получает также и по одному из частичных секретов для каждого из сообщений, которые не использовались при шифровании.

Поскольку сервер знает *s*, то он может восстановить ε дважды и косвенно убедиться, что эти сообщения сформированы тем, кто тоже знает s и ε. А поскольку в каждом сеансе счетчик меняется, то повторно использовать сообщения предыдущего сеанса не выйдет. Авторы в работе [1] предложили два протокола аутентификации, отличающиеся лишь тем, что во втором протоколе вместо общего пароля *key* клиент и сервер хранят частичные секреты key_0 и key_1, соответственно, полученные путем применения (2, 2)-пороговой схемы Шамира к *key*. Из-за ограничений по объему, приведем лишь первый из этих протоколов. В нем через *Recover*() обозначена функция восстановления секрета по схеме Шамира; ε_i — случайное число, выбираемое клиентом для подтверждения знания общего с сервером секрета $s = key + ctr$; $\left(\varepsilon_i^0, \varepsilon_i^1\right)$ — пара частичных секретов, полученных путем применения схемы Шамира к ε_i; $\varepsilon_{i_0}, \varepsilon_{i_1} \in \mathbf{Z}_p^*$ — два случайных числа, произведение которых равно ε_i (это и есть мультипликативная схема полного согласия). Через '/' обозначена операция деления (умножения на обратный элемент в Z_p), '||' – символ конкатенации (склейки) слов.

Протокол аутентификации TUS-1

Клиент		**Сервер**
		Проверяет, что $d_i / b_i \neq 0, d_i / \varepsilon_{i_1} \neq 0;$
$\varepsilon_i, b_i, d_i \xleftarrow{R} \mathbf{Z}_p^*;$		$\frac{d_i}{\varepsilon_i} = \frac{d_i}{\varepsilon_{i_1}} \times \frac{1}{\varepsilon_{i_0}};$
$\varepsilon_i = \varepsilon_{i_0} \cdot \varepsilon_{i_1} = Recover\left(\varepsilon_i^0, \varepsilon_i^1\right);$		
$M = \varepsilon_{i_0} \,\|\| \, \varepsilon_i^0 \,\|\| \, \varepsilon_i^1 \left(key + ctr\right) \,\|\|$	$\xrightarrow{M}$	$d_i A_0 = b_i \left(key + ctr\right) \cdot \frac{d_i}{b_i} - \left(key + ctr\right) \cdot \frac{d_i}{\varepsilon_i} \cdot \varepsilon_i^0;$
$\|\| \, b_i \left(key + ctr\right) \,\|\| \, d_i / b_i \,\|\| \, d_i / \varepsilon_{i_1}.$		$d_i A_1 = b_i \left(key + ctr\right) \cdot \frac{d_i}{b_i} - \left(key + ctr\right) \cdot \frac{d_i}{\varepsilon_i} \cdot \varepsilon_i^1;$
		Проверяет, что $Recover(d_i A_0, d_i A_1) = 0.$

Во-первых, отметим, что действия на стороне сервера могли бы быть гораздо проще — ему надо было лишь из сообщения *M* найти, расшифровав соответствующие части *M*, два оставшихся частичных секрета ε_i^1 и ε_{i_1}, а затем восстановить ε_i по соответствующим формулам и сравнить. С этой точки зрения, во второй части сообщения *M* клиент пересылает бесполезные числа — вместо этого можно было бы просто отправить $\varepsilon_{i_1}\left(key + ctr\right)$. Во-вторых, хоть авторы и «доказали» теоретико-информационную стойкость протокола TUS-1, однако они сделали это не достаточно строго, что позволило нам провести успешную атаку, раскрывающую значение секрета ($key + ctr$). Аналогичной слабостью обладает и второй протокол TUS-2. Этот факт говорит, во-первых, о том, насколько важно использовать формальные методы доказательства теорем, а во-вторых, не достаточную компетентность рецензентов, казалось бы, издания с высоким рейтингом и квартилем Q1.

Таким образом, предложенные совсем недавно известными учеными Токийского научного университета протоколы оказались не стойкими, хотя утверждалась их теоретико-информационная стойкость. В докладе мы также обсудим возможность построения теоретико-информационно стойких протоколов односторонней аутентификации в рассмотренных условиях.

Список использованных источников:

1. K. Iwamura and A. A. A. M. Kamal, "Secure User Authentication With Information Theoretic Security Using Secret Sharing-Based Secure Computation," in *IEEE Access*, vol. 13, pp. 9015-9031, 2025, doi: 10.1109/ACCESS.2025.3526632
2. A. A. A. Mohd Kamal and K. Iwamura, “Conditionally secure multiparty computation using secret sharing scheme for n <2k-1 (Short Paper),” in *Proc. 15th Annu. Conf. Privacy, Secur. Trust (PST)*, Aug. 2017, pp. 225–2255.
3. K. Iwamura and A. A. A. M. Kamal, “Communication-efficient secure computation of encrypted inputs using (k, n) threshold secret sharing,” IEEE Access, vol. 11, pp. 51166–51184, 2023.
4. A. Shamir, “How to share a secret,” Commun. ACM, vol. 22, no. 11, pp. 612–613, Nov. 1979.

Федорова А.А., Полтавцева М.А.
Санкт-Петербургский политехнический университет Петра Великого, Санкт-Петербург

ШИФРОВАНИЕ В БАЗАХ ДАННЫХ NOSQL ДЛЯ ЗАЩИТЫ ОТ ПРИВИЛЕГИРОВАННОГО НАРУШИТЕЛЯ

Развитие NoSQL-систем, ориентированных на хранение больших объёмов слабо структурированных данных, делает особенно актуальной задачу защиты информации в условиях недоверенной среды. Стандартные серверные механизмы защиты оказываются недостаточными, если рассматривать угрозу привилегированного нарушителя, то есть администратора или иного субъекта, имеющего расширенный доступ к инфраструктуре хранения. В этом случае возникает необходимость не просто зашифровать данные, а понять, каким образом можно сохранить работоспособность и производительность и запросов к базе данных и одновременно не допустить раскрытия информации на стороне сервера. Именно эти аспекты определяют актуальность настоящего исследования.

В работе рассматривается задача построения архитектуры для защиты данных в MongoDB от привилегированного нарушителя с использованием модели луковичного (многослойного) шифрования [1]. Существующие работы в данной области показывают, что поддержка операций над зашифрованными данными достигается ценой появления утечек различного типа: утечек равенства, порядка, других косвенных признаков [1,3]. Это означает, что задача не сводится к простому усилению криптографической стойкости: необходимо определить, какие утечки являются допустимыми для конкретной операции.

Существующие работы в данной области опираются на концепцию клиентской обработки запросов [1, 2], однако в современных исследованиях [3] выделяют два принципиально разных подхода к организации криптографических слоёв:

оследовательная (иерархическая) архитектура. Слои шифрования накладываются друг на друга. Главный недостаток подхода – нарушение принципа минимального раскрытия: для выполнения операции на внутреннем слое необходимо расшифровать все внешние, что приводит к каскадному ослаблению криптостойкости и суммированию утечек информации.

араллельная (горизонтальная) архитектура. Для каждого типа защищаемых данных создаются независимые колонки, каждая из которых зашифрована своим алгоритмом. Это позволяет выполнять запросы непосредственно на нужном слое, полностью изолируя утечки и предотвращая атаки на основе композиции утечек.

В рамках данной работы для реализации выбрана параллельная архитектура.

Далее был произведен анализ допустимых операций базы данных MongoDB, были выделены три ключевые группы операций над значениями, требующие независимой

криптографической поддержки: проверки на точное равенство ($eq, $in), операции сравнения и диапазонного поиска ($gt, $lt, $gte, $lte) и арифметические агрегации (сумма, среднее).

Каждому слою в архитектуре луковичного шифрования соответствует свой допустимый профиль утечек, который является неизбежным компромиссом для обеспечения работы серверных операций. В связи с этим, выбор конкретных криптографических алгоритмов для реализации слоёв шифрования обусловлен соответствием присущих им уровней утечек тем требованиям, которые предъявляет логика слоя шифрования луковицы.

В соответствии с этими группами была разработана трехслойная параллельная структура шифрования для *числовых* данных:

Слой DET – для поддержки операций точного равенства и поиска по ключу.

Слой HOPE – для безопасного выполнения операций сравнения без раскрытия интервалов между значениями.

Слой HOM – для выполнения гомоморфных агрегаций.

Для *строковых* данных применяется упрощенная конфигурация, ограниченная слоями DET и HOPE.

Для слоя DET (детерминированное шифрование) был выбран режим AES-GCM-Шифрование в режиме счетчика над полем Галуа с синтетической синхропосылкой). Основное требование к DET-слою – поддержка операций точного равенства (например, $eq, $in, $lookup в MongoDB). Для выполнения этой задачи алгоритм обязан при каждом вызове преобразовывать одинаковые открытые тексты в идентичные шифртексты. Режим AES-GCM-SIV идеально подходит для этой роли: в спецификации этого алгоритма прямо указано, что подобный режим работы приводит к утечке информации о равенстве открытых текстов (equality leakage). В контексте архитектуры проектируемой системы эта утечка является осознанным допущением, присутствующим в системе в рамках DET-слоя по теоретическому построению. Именно такая концепция позволяет серверу MongoDB строить индексы и выполнять поиск по точным совпадениям без расшифрования данных.

Для слоев HOM (гомоморфные агрегации) и HOPE (сравнение с сохранением порядка) была выбрана криптосистема Пэйе (Paillier). Выбор обоснован тем, что её профиль утечек строго ограничен её функциональными свойствами. Она раскрывает ровно ту информацию, которая извлекается в результате разрешенных вычислений (например, сам факт выполнения сложения для HOM-слоя, или только знак разности при использовании специального ключа сравнения в HOPE-слое) [4].

В результате проведенного исследования были сформулированы ключевые требования к итоговой системе:

золяция утечек: запросы должны маршрутизироваться строго на тот криптографический слой, который минимально необходим для их выполнения.

езависимость ключей: параллельные слои должны использовать независимые криптографические ключи.

розрачная интеграция: криптографические трансформации должны осуществляться "на лету" с помощью прокси-сервера, перехватывающего wire-сообщения MongoDB, без внесения изменений в ядро самой СУБД.

В результате проведённого исследования были сформулированы требования к структуре слоёв. Дальнейшая работа связана с программной реализацией описанной системы луковичного шифрования и проведением экспериментов по оценке её производительности.

Библиографический список:

1. Popa R. A., Redfield C. M. S., Zeldovich N., Balakrishnan H. CryptDB: Protecting Confidentiality with Encrypted Query Processing // Proceedings of the 23rd ACM Symposium on Operating Systems Principles (SOSP). — 2011. — P. 85–100.
2. Popa R. A., Redfield C. M. S., Zeldovich N., Balakrishnan H. CryptDB: Processing Queries on an Encrypted Database // Communications of the ACM. — 2012. — Vol. 55, No. 9. — P. 103–111. — DOI: 10.1145/2330667.2330691.
3. Wang X., Li Y., Zhang H. A Survey of Data Protection Methods in Cloud DBMS [Электронный ресурс] // arXiv. — 2024. — URL: https://arxiv.org/abs/2411.17009.
4. Chen Y., et al. HOPE: Homomorphic Order-Preserving Encryption for Outsourced Databases – A Stateless Approach // IEEE Transactions on Dependable and Secure Computing. — 2024. — Vol. 21, № 3. — P. 1450–1464.

6. ПОДГОТОВКА КАДРОВ В СФЕРЕ ИНФОРМАЦИОННОЙ БЕЗОПАСНОСТИ

Ахрамеева К.А.
ООО «Газинформсервис», Санкт-Петербург

КИБЕРПОЛИГОН КАК ИНСТРУМЕНТ ПРАКТИЧЕСКОЙ ОТРАБОТКИ КИБЕРИНЦИДЕНТОВ В СИСТЕМЕ ПОДГОТОВКИ КАДРОВ ПО ИНФОРМАЦИОННОЙ БЕЗОПАСНОСТИ

Аннотация: В условиях дефицита квалифицированных кадров в сфере информационной безопасности особенно остро стоит вопрос практической подготовки выпускников вузов. В тезисах рассматривается роль киберполигонов как инструмента отработки киберинцидентов в учебном процессе. Анализируются основные проблемы взаимодействия бизнеса и образования, а также предлагаются подходы к интеграции практико-ориентированных методов обучения..

Ключевые слова: киберинцидент, киберарена, практико-ориентированное обучение, информационная безопасность.

Дефицит квалифицированных кадров в сфере информационной безопасности остаётся одной из острых проблем российского рынка. Количество выпускников растет, однако качество их подготовки зачастую не соответствует реальным задачам бизнеса. Как отмечается в работе [1], одной из ключевых причин кадрового дефицита является недостаточная адаптивность вузовских образовательных программ к быстро меняющимся требованиям рынка труда: учебные планы вузов остаются преимущественно теоретически ориентированными и не предусматривают достаточного объёма практических занятий, связанных с реальными инцидентами и моделированием кибератак. В результате выпускники не обладают необходимыми компетенциями для эффективного решения актуальных задач информационной безопасности, и работодателям приходится дополнительно обучать молодых специалистов. В то же время, согласно исследованию, в 2025 году с нехваткой ИБ-специалистов сталкивались 71% опрошенных организаций [2], что подтверждает актуальность обозначенной проблемы.

Для качественной подготовки специалистов крайне важно, чтобы студенты имели возможность проходить практику на актуальном программном обеспечении, используемом коммерческими и государственными организациями в реальной профессиональной деятельности. В качестве такого инструмента выступает киберарена – виртуальная безопасная среда, специально созданная для формирования и оценки практических навыков обучающихся. Ключевое удобство киберарены заключается в том, что она позволяет без риска для реальных систем разворачивать различные инфраструктуры, устанавливать современные средства защиты информации и инструменты мониторинга. Студенты работают с тем же программным обеспечением, с которым им предстоит сталкиваться в будущей работе, а преподаватели получают готовые сценарии занятий, основанные на

реальных инцидентах, что значительно снижает порог входа для организации качественной практической подготовки.

Киберарена позволяет моделировать отраслевые инфраструктуры – нефтегазовый сектор, финансовый сектор, научные организации и другие, что даёт возможность готовить специалистов с учётом специфики конкретных секторов экономики. Среда поддерживает проведение киберучений с оценкой входного и целевого уровней подготовки, обеспечивая объективный контроль прогресса обучающихся. На её базе могут проводиться демо-экзамены и итоговые аттестации, что повышает практическую значимость образовательного процесса. Таким образом, киберарена выступает не только как технологический, но и как организационно-методический инструмент сближения академического образования и реальных потребностей рынка труда в сфере информационной безопасности. Её внедрение позволяет системно повысить качество подготовки студентов, готовых к эффективному реагированию на киберинциденты в реальных условиях.

Список литературы:

храмеева, К. А. Проблематика кадрового дефицита в информационной безопасности в России: вызовы и опыт практико-ориентированной подготовки специалистов / К. А. Ахрамеева // Информационная безопасность регионов России (ИБРР-2025) : Материалы XIV Санкт-Петербургской межрегиональной конференции, Санкт-Петербург, 29–31 октября 2025 года. – Санкт-Петербург: Санкт-Петербургское Общество информатики, вычислительной техники, систем связи и управления, 2025. – С. 394-395.

ёрчИнформ. 71% компаний РФ испытывают нехватку кадров на рынке

Белов Е.Б.

Федеральное учебно-методическое объединение в системе высшего образования по УГСН 10.00.00 «Информационная безопасность»

ПОДХОДЫ К ОПРЕДЕЛЕНИЮ ЦЕЛЕВЫХ УСТАНОВОК СОЗДАНИЯ И ФУНКЦИОНИРОВАНИЯ ПИЛОТНЫХ МНОГОФУНКЦИОНАЛЬНЫХ УЧЕБНО-НАУЧНЫХ (ПРОИЗВОДСТВЕННЫХ) ЦЕНТРОВ ПО ПРОБЛЕМАМ ИНФОРМАЦИОННОЙ БЕЗОПАСНОСТИ В ФЕДЕРАЛЬНЫХ ОКРУГАХ

В статье рассмотрены вопросы создания пилотных многофункциональных учебно-научных (производственных) центров по проблемам информационной безопасности в Центральном и Сибирском федеральных округах, предложены целевые установки и задачи деятельности указанных центров.

Ключевые слова. *Информационная безопасность, Концепция создания системы многофункциональных учебно-научных (производственных) центров, цели, задачи пилотных центров, система подготовки специалистов, федеральный государственный образовательный стандарт высшего образования.*

Во исполнение Поручения Президента Российской Федерации от 6 декабря 2022 г. № Пр-2330 «О совершенствовании подготовки кадров для обеспечения информационной безопасности Российской Федерации» Минобрнауки России с участием ФУМО ВО ИБ проводит работу по созданию и запуску пилотных многофункциональных учебно-научных (производственных) центров по проблемам информационной безопасности в Центральном

и Сибирском федеральных округах (далее – ИБ, пилотные УНЦ ИБ, ЦФО, СФО). Данная работа проводится на основании разработанной с участием ФУМО ВО ИБ Концепции создания системы многофункциональных учебно-научных (производственных) центров по проблемам информационной безопасности в федеральных округах на базе ведущих образовательных организаций высшего образования, утвержденной заместителем Министра науки и высшего образования Российской Федерации А.В. Омельчуком 22 ноября 2024 г. (далее соответственно – Концепция).

В ходе открытого конкурса, проведенного Минобрнауки России совместно с ФУМО ВО ИБ, принято решение о создании пилотных УНЦ ИБ на базе РТУ – МИРЭА, МГТУ им. Н.Э. Баумана и ТУСУР.

Создаваемые пилотные УНЦ ИБ выступают ключевыми элементами региональных систем подготовки специалистов и развития технологий информационной безопасности. Центры ориентированы на интеграцию образовательных, научных и производственных возможностей ЦФО и СФО, обеспечивают комплексный подход к формированию компетенций, развитию научного потенциала и внедрению передовых решений в области информационной безопасности.

Создание пилотных УНЦ ИБ СФО и ЦФО является ключевым этапом реализации Поручения Президента Российской Федерации, направленного на формирование распределённой национальной сети учебно-научных структур в области информационной безопасности. Пилотные центры выполняют функцию исходных площадок, на которых в первоочередном порядке апробируются модели взаимодействия образовательных и научных организаций, государственных органов и предприятий, а также формируются подходы к организации лабораторной учебно-научной инфраструктуры и разрабатываются методические и технологические решения.

Накопленный в ходе их деятельности практический опыт предназначается для дальнейшего использования при развертывании полной сети из восьми УНЦ ИБ во всех федеральных округах Российской Федерации.

Пилотный УНЦ ИБ призван обеспечить полный цикл формирования инновационных организационных, образовательных и технологических решений, включающий их разработку, последующее тестирование внутри центров и во взаимодействии с образовательными, научно-производственными организациями, органами власти в соответствующих федеральных округах, а также дальнейшее совершенствование этих решений на основе полученных результатов.

Одной из ключевых функций пилотных центров является внедрение и апробация модели центров коллективного пользования (далее - ЦКП). Такая модель предполагает предоставление лабораторий, полигонов и высокотехнологичного оборудования широкому кругу образовательных организаций высшего образования и научных организаций для формирования практических навыков и компетенций обучающихся, а также для реализации сетевых образовательных программ, проведения передовых научных исследований.

Пилотные центры призваны сформировать устойчивую и технологически оснащённую инфраструктурную базу ЦКП, определить требования к оснащению лабораторий, разработать типовые модули практико-ориентированной подготовки и механизмы совместного доступа к оборудованию.

Не менее важной задачей пилотных УНЦ ИБ является отработка механизмов межведомственной координации и управления. В соответствии с Концепцией, деятельность центров координируется Минобрнауки России с участием ФУМО ВО ИБ, ФСБ России и ФСТЭК России, которые участвуют в определении требований к составу и техническому оснащению входящих в центры лабораторий и полигонов.

Пилотные УНЦ ИБ предполагают три ключевых направления деятельности: учебно-методическое, научно-исследовательское и производственное, а также информационно-коммуникационное. Выделение трёх ключевых направлений деятельности обеспечивает комплексное сочетание подготовки высококвалифицированных специалистов, проведения

передовых научных исследований и эффективного распространения знаний и технологий в сфере кибербезопасности.

Учебно-методическое направление УНЦ ИБ ориентировано на совершенствование подготовки специалистов в сфере информационной безопасности в регионе с перспективой тиражирования успешных практик на другие территории. Центр концентрирует компетенции по ключевым направлениям подготовки дипломированных специалистов в области ИБ, включая безопасность компьютерных систем и сетей, автоматизированных систем, телекоммуникационных систем и сетей, финансово-экономических структур, систем и механизмов искусственного интеллекта, а также методы и технологии защиты информации.

Методическая работа УНЦ ИБ в значительной степени направлена на совершенствование технического и методического обеспечения образовательного процесса, а также на разработку типовых программных документов для перспективных федеральных государственных образовательных стандартов. Особое внимание уделяется дополнительному профессиональному образованию в сфере ИБ, включая повышение квалификации и профессиональную переподготовку специалистов с различной базовой подготовкой, работающих в разных отраслях, что позволяет пилотному центру развивать и укреплять региональный кадровый потенциал в области информационной безопасности.

Научно-исследовательское и производственное направление будет обеспечивать проведение фундаментальных и прикладных исследований, а также разработку отечественных технологий защиты информации. В рамках этого направления предлагается реализация проектов по разработке и внедрению технологий квантовой коммуникации, семантической фильтрации сетевого трафика, комплексной защите информационных каналов, технологиям доверенного взаимодействия пользователей в цифровой среде.

Информационно-коммуникационное направление формирует единое информационно-аналитическое и коммуникационное пространство в сфере информационной безопасности. Направление состоит из двух блоков:

1. Коммуникационный и координационный блок, который включает: определение правил, норм и создание площадок коммуникации стейкхолдеров; создание основы эффективного взаимодействия государственных органов, академических и профессиональных сообществ, производственных предприятий сферы ИБ; участие в разработке проектов локальных нормативно-правовых актов и распорядительных документов в предметной области; предоставление информационной и методической поддержки, а также доступа к ресурсам пилотного УНЦ ИБ.
2. Информационно-просветительский блок, который включает: информирование заинтересованных лиц о предоставляемых пилотным УНЦ ИБ возможностях и услугах; сопровождение и освещение профориентационных, просветительских, образовательных, научно-методических мероприятий, организация и проведение окружных киберучений, межвузовских практических тренировок и отраслевых учений, студенческих олимпиад и конкурсов в области информационной безопасности, организуемых на базе пилотного УНЦ ИБ (в том числе совместно с партнерами).

Наиболее важным в деятельности пилотных центров является учебно-методическое направление деятельности, в рамках которого должны решаться следующие задачи:

- подготовка предложений по совершенствованию системы подготовки специалистов в области ИБ по основным образовательным программам (далее – ООП ИБ) на основе анализа потребности в кадрах, координации деятельности и обмена опытом образовательных организаций федерального округа с учетом развития ИТ и противодействия новым угрозам безопасности информации;
- координация усилий образовательных организаций федерального округа в совершенствовании научно-методического обеспечения образовательного процесса по реализации ООП ИБ, включая разработку современных методик обучения и педагогических технологий, учебников, учебных и учебно-методических пособий,

электронных образовательных курсов, открытых онлайн-курсов, лабораторных практикумов с применением отечественного оборудования, фондов оценочных средств, в том числе путем создания общих информационно-методических ресурсов;

- развитие практико-ориентированной системы подготовки специалистов, которое осуществляется с ориентацией на формирование профессиональных компетенций, востребованных отраслью и регуляторами, и включает актуализацию содержания образовательных программ, усиление практической подготовки и интеграцию результатов научно-исследовательской деятельности в рамках актуальных специализаций;
- организация дополнительного профессионального образования, обеспечивающего для органов власти, организаций и предприятий федерального округа систему поддержания и непрерывного развития профессиональных компетенций специалистов в сфере ИБ, в том числе на основе постоянного взаимодействия с регуляторами в области ИБ (ФСТЭК России и ФСБ России) и руководителями предприятий и организаций округа (заказчиками на подготовку кадров);
- повышение квалификации и профессиональная переподготовка преподавателей вузов федеральных округов на базе пилотных УНЦ ИБ по перспективным образовательным программам;
- совершенствование материально-технического (лабораторного) обеспечения образовательного процесса по реализации образовательными организациями федерального округа образовательных программ в сфере ИБ на основе разработки и создания современных киберполигонов и учебно-тренировочных средств отечественных производителей, включая создание комплексного окружного учебно-научного полигона по проблемам ИБ, в первую очередь, в области защиты информации значимых объектов критической информационной инфраструктуры Российской Федерации.

С учетом изложенного, предлагается следующий примерный перечень работ по развитию учебно-методического направления пилотного УНЦ ИБ:

– разработка модели сетевого взаимодействия по образовательным программам с образовательными организациями;

– организация лабораторий ЦКП ИБ для реализации образовательных программ, в т.ч. в сетевой форме;

– разработка образовательных модулей для освоения общепрофессиональных компетенций;

– разработка образовательных модулей для освоения профессиональных компетенций и проведения научно-исследовательской работы студентами;

– реализация образовательных модулей в программах дополнительного образования научно-педагогических работников организаций высшего образования;

– разработка и внедрение образовательных модулей по профильным дисциплинам в области ИБ с использованием электронного обучения в рамках сетевых образовательных программ.

Таким образом, основной задачей пилотных УНЦ ИБ является апробация организационных, образовательных и инфраструктурных моделей подготовки, переподготовки и повышения квалификации кадров в области информационной безопасности, развитие преподавательских компетенций образовательных организаций федерального округа для последующего тиражирования и применения в рамках формирования распределенной системы УНЦ ИБ.

Список литературы:

1. Федеральный закон от 29 декабря 2012 г. N 273-ФЗ "Об образовании в Российской Федерации" (с изменениями и дополнениями). – Текст: непосредственный.
2. Концепция создания системы многофункциональных учебно-научных (производственных) центров по проблемам информационной безопасности в федеральных округах на базе ведущих образовательных организаций высшего

образования, утвержденная заместителем Министра науки и высшего образования Российской Федерации А.В. Омельчуком 22 ноября 2024 г. – Текст: непосредственный.

Белов Е.Б.
Федеральное учебно-методическое объединение в системе высшего образования по УГСН 10.00.00 «Информационная безопасность»

ПОДХОДЫ К ФОРМИРОВАНИЮ КЛЮЧЕВЫХ НАПРАВЛЕНИЙ СОВЕРШЕНСТВОВАНИЯ СИСТЕМЫ ПОДГОТОВКИ КАДРОВ С ВЫСШИМ ОБРАЗОВАНИЕМ В ОБЛАСТИ ОБЕСПЕЧЕНИЯ ИНФОРМАЦИОННОЙ БЕЗОПАСНОСТИ РОССИЙСКОЙ ФЕДЕРАЦИИ

В статье рассмотрены результаты реализации Концепции развития кадрового обеспечения в области информационной безопасности в Российской Федерации на долгосрочную перспективу до 2025 года, состояние и факторы, влияющие на систему подготовки специалистов с высшим образованием в области информационной безопасности, выработаны ключевые направления по ее совершенствованию.

Ключевые слова. Информационная безопасность, Концепция развития кадрового обеспечения в области информационной безопасности, направления совершенствования системы подготовки кадров с высшим образованием, федеральный государственный образовательный стандарт высшего образования.

В соответствии со Стратегией национальной безопасности Российской Федерации, Доктриной информационной безопасности Российской Федерации, Указами Президента Российской Федерации «О национальных целях развития Российской Федерации на период до 2030 года и на перспективу до 2036 года», «О дополнительных мерах по обеспечению информационной безопасности Российской Федерации», а также решениями Совета Безопасности Российской Федерации одним из приоритетных направлений государственной политики по развитию кадрового обеспечения информационной безопасности (далее – ИБ) является совершенствование подготовки кадров с высшим образованием в области обеспечения информационной безопасности Российской Федерации.

В период 2015 – 2025 гг. основным документом стратегического планирования в кадровом обеспечении ИБ является **Концепция (далее – Концепция) развития кадрового обеспечения в области информационной безопасности в Российской Федерации на долгосрочную перспективу** (утв. Решением коллегии Военно-промышленной комиссии Российской Федерации от 9 марта 2017 г. № ВПК-бр).

По оценке экспертов Аппарата Совета Безопасности Российской Федерации, Минобрнауки России и ФУМО ВО ИБ положения Концепции, в части конкретных задач по совершенствованию высшего образования, в основном выполнены.

Минобрнауки России постоянно увеличивает объем контрольных цифр приема по УГСН «Информационная безопасность». Разработаны ФГОС ВО по специальностям и направлениям в области информационной безопасности, единственные в системе высшего образования, имеющие расширенные требования к кадровому, учебно-методическому и материально-техническому обеспечению учебного процесса.

Разработано методическое обеспечение реализации всех ФГОС ВО, и основных образовательных программ в области информационной безопасности. Минобрнауки

России проведены мероприятия по созданию пилотной зоны многофункциональных учебно-научных (производственных) центров по проблемам информационной безопасности.

Создан и работает механизм целевого обучения специалистов по защите информации по образовательным программам высшего образования в интересах отдельных государственных органов и организаций оборонно-промышленного комплекса.

В рамках федерального проекта «Информационная безопасность» национальной программы «Цифровая экономика Российской Федерации» в течение 2020 – 2023 гг. проведен комплекс мероприятий: по повышению квалификации педагогических работников; по грантовой поддержке аспирантов, соискателей и молодых ученых проводящих исследования по перспективным направлениям в области информационной безопасности; по проведению олимпиад, киберучений и иных интеллектуальных соревнований в области информационной безопасности со студентами образовательных организаций.

Вместе с тем, ряд задач в полном объеме реализовать не удалось. В первую очередь это касается инфраструктурного, ресурсного и кадрового развития подразделений образовательных организаций высшего образования, реализующих программы в области ИБ.

Исследования, проведенные экспертами ФУМО ВО ИБ и анализ предложений работодателей, позволяют определить следующие дополнительные уязвимости в системе высшего образования в области ИБ.

Доля преподавателей профильных кафедр по ИБ, имеющих базовое образование или прошедших профессиональную переподготовку в указанной области, составляет не более 50 процентов. Сохраняется дефицит преподавателей высшей квалификации по научным специальностям в области информационной безопасности.

Не в полной мере развита кооперация образовательных организаций и организаций-работодателей в проведении практик обучающихся на базе профильных организаций в области информационной безопасности.

Специалисты-практики недостаточно привлекаются к участию в образовательном процессе. Задача по стимулированию и закреплению педагогических кадров в области ИБ требует дополнительного решения.

Используемая технологическая и методическая база реализации образовательных программ в области информационной безопасности не в полной мере отвечает современному научно-техническому уровню информационных технологий, современным способам и средствам обеспечения защиты информации. Отсутствие специализированных учебно-тренировочных и лабораторных комплексов оказывает негативное влияние на качество формирования практико-ориентированных компетенций у выпускников.

Проблемным вопросом остается несовершенство существующих ФГОС с учетом возрастания требований к результатам обучения и условиям реализации основных образовательных программ в области ИБ.

Уровень участия образовательных организаций в научных разработках в области ИБ является низким. Таким образом, в настоящее время состояние системы подготовки специалистов с высшим образованием в области информационной безопасности не в полной мере отвечает потребностям обеспечения заинтересованных государственных органов и организаций квалифицированными кадрами.

С учетом изложенного, а также реализации требований уполномоченных государственных органов по достижению качества подготовки специалистов по защите информации, которые должны уверенно противодействовать новым вызовам и угрозам в информационной сфере, обуславливают разработку ключевых направлений совершенствования системы подготовки специалистов по защите информации с высшим

образованием.

Важным условием совершенствования системы подготовки специалистов по защите информации с высшим образованием является проектирование новой модели такой подготовки в рамках формирования общенациональной системы высшего образования.

Комплексный анализ состояния системы подготовки специалистов по защите информации с высшим образованием и факторов, влияющих на ее развитие, позволил выработать следующие ключевые направления совершенствования данной системы:

– формирование новой модели отдельной укрупненной группы специальностей «Информационная безопасность» в рамках формирования национальной системы высшего образования, направленной на реализацию стратегического национального приоритета «Информационная безопасность»;

– разработка ФГОС ВО нового поколения и образовательных программ в области информационной безопасности в рамках формирования национальной системы высшего образования с учетом утвержденных (актуализированных) профессиональных стандартов и требований нормативных правовых актов в области информационной безопасности, сочетающих фундаментальность, междисциплинарность и практико-ориентированность обучения;

– разработка и реализация инновационных опережающих образовательных программ во взаимодействии с работодателями для подготовки специалистов, способных обеспечить нейтрализацию угроз, в том числе, неизвестных на период получения образования;

– разработка и внедрение во все ФГОС высшего образования нового поколения компетенции (образовательных модулей) по информационной безопасности для включения в основные образовательные программы высшего образования;

– разработка мер по усилению государственного контроля качества образования при реализации основных профессиональных образовательных программ высшего образования в области информационной безопасности;

– создание системы многофункциональных учебно-научных (производственных) центров по проблемам информационной безопасности в федеральных округах на базе ведущих образовательных организаций высшего образования, подведомственных Минобрнауки России;

– опережающее обеспечение специализированными высокотехнологичными комплексами учебно-тренировочных средств, компьютерными полигонами, средствами защиты информации российских производителей и базовыми лабораторными практикумами структурных подразделений образовательных организаций;

– создание головного федерального учебно-методического центра по сопровождению деятельности системы многофункциональных учебно-научных (производственных) центров и межрегиональных центров компетенций по проблемам в области ИБ на базе образовательной организации, обеспечивающей деятельность ФУМО ВО ИБ;

– совершенствование мер по повышению квалификации профессорско-преподавательского состава и задействованию учебных и производственных практик при реализации основных профессиональных образовательных программ высшего образования в области ИБ, предусмотрев внесение в нормативные правовые акты изменений по переводу в третью стоимостную группу специальностей и направлений подготовки по государственным услугам по реализации данных образовательных программ;

– совершенствование мер и механизмов материального стимулирования руководителей и научно-педагогических работников кафедр (структурных подразделений) образовательных организаций высшего образования, реализующих основные образовательные программы в области ИБ;

– совершенствование мер по проведению профессиональной стажировки педагогических работников, реализующих образовательные программы в области ИБ, в организациях, занимающихся разработкой средств защиты информации или оказывающих услуги в данной сфере;

– проведение с мерами государственной поддержки олимпиад, учений, конкурсов научных работ и иных интеллектуальных соревнований федерального и регионального уровня в области ИБ со студентами и школьниками;

– грантовая поддержка аспирантов, соискателей и молодых ученых, проводящих исследования по перспективным направлениям в области ИБ;

– развитие аспирантуры и докторантуры, формируемых на базе авторитетных научных школ образовательных и научных организаций, ведущих исследования в области ИБ;

– разработка и внедрение виртуальных лабораторных сред для формирования практико-ориентированных компетенций по нейтрализации внутренних и внешних угроз информационной безопасности с участием отечественных разработчиков средств защиты информации;

– управление жизненным циклом образовательной программы во взаимодействии с заказчиками–организациями реального сектора экономики;

– формирование универсальных компетенций (системное мышление, коммуникации, принятие решений, командная работа, навыки самообучения, компетенции цифровой экономики), активизация патриотической, профориентационной и воспитательной работы в рамках современных требований к интеграции образовательной деятельности
и воспитательного процесса.

Таким образом, при формировании проекта новой редакции Концепции подготовки, профессиональной переподготовки и повышения квалификации кадров в области обеспечения информационной безопасности Российской Федерации на период до 2035 года предлагается учесть вышеперечисленные направления совершенствования системы подготовки специалистов с высшим образованием.

В целях достижения реальных результатов предлагается разработать основные мероприятия по реализации Концепции, которые должны учитываться в Национальном плане обеспечения информационной безопасности, разрабатываемом участниками национальной системы обеспечения информационной безопасности.

Список литературы:

1. Федеральный закон от 29 декабря 2012 г. N 273-ФЗ "Об образовании в Российской Федерации" (с изменениями и дополнениями). – Текст: непосредственный.
2. Указ Президента Российской Федерации от 12.05.23 № 343 «О пилотном проекте, направленном на изменение уровней профессионального образования».
3. Научно-методические материалы Всероссийской научно-практической конференции «Кадровое обеспечение информационной безопасности Российской Федерации» и Пленума ФУМО ВО ИБ. Проект ФГОС ВО-4 по УГСН 34 «Информационная безопасность»/ Белов Е.Б., Райх В.В. и др. // Москва, 16–18 апреля 2025: М, 2025 г. – 93 с. – Текст: непосредственный.

Сторожик В.С.
Арктический и антарктический научно-исследовательский институт, Санкт-Петербург

НОРМАТИВНО-ПРАВОВОЕ И МЕТОДИЧЕСКОЕ ОБЕСПЕЧЕНИЕ ГОСУДАРСТВЕННОЙ ПОЛИТИКИ В ОБЛАСТИ ТЕХНИЧЕСКОЙ ЗАЩИТЫ ИНФОРМАЦИОННЫХ РЕСУРСОВ РОССИЙСКОЙ ФЕДЕРАЦИИ

В Стратегии национальной безопасности Российской Федерации к национальным интересам отнесено развитие безопасного информационного пространства, а среди стратегических национальных приоритетов, направленных на обеспечение и защиту национальных интересов Российской Федерации, выделена информационная безопасность (ИБ) [1].

В Доктрине информационной безопасности Российской Федерации приведена организационная основа системы обеспечения ИБ, указан состав её участников, а также принципы деятельности государственных органов по обеспечению ИБ, их задачи по обеспечению ИБ, развитию и совершенствованию системы

В целях дальнейшего повышения устойчивости и безопасности функционирования информационных ресурсов Российской Федерации в органах и организациях на заместителя руководителя возложены полномочия по обеспечению ИБ, а также созданы структурные подразделения по обеспечению ИБ [3, 4].

В проекте Доктрины информационной безопасности Российской Федерации выделены основные принципы достижения цели обеспечения ИБ: необходимости и достаточности сил и средств обеспечения ИБ, комплексности мер по обеспечению ИБ и непрерывности их осуществления, а также неразрывности обеспечения ИБ с проведением всех видов работ, связанных с созданием и эксплуатацией объектов информационной инфраструктуры [5].

С 1 марта 2026 г. введены в действие Требования о защите информации, содержащейся в государственных информационных системах, иных информационных системах государственных органов, государственных унитарных предприятий, государственных учреждений (Требования) [6], которые заменяют Требования о защите информации, не составляющей государственную тайну, содержащейся в государственных информационных системах (ГИС) [7].

Требования значительно расширяют область регулирования защиты информации:

независимо от характера обрабатываемой информации (ограниченного доступа эта информация или нет), кроме ГИС Требования распространяются также на иные информационные системы (ИС) государственных органов, государственных унитарных предприятий, государственных учреждений и муниципальные ИС, если иное не установлено законодательством Российской Федерации;

устанавливается, что в случае передачи из ГИС информации ограниченного доступа в негосударственную ИС, уровень её защищенности должен устанавливаться соответствующим государственным заказчиком (оператором ГИС) в соответствии с Требованиями.

Для достижения целей защиты информации (ЗИ) должно проводиться 21 мероприятие по ЗИ, а также в ИС должно быть реализовано 17 групп мер ЗИ, которые регламентированы в рамках полномочий ФСТЭК России [8].

Во исполнение Требований в государственных органах и организациях необходимо разработать планы перехода, содержащие описание мероприятий и мер по ЗИ, со сроками их реализации [9].

Также ФСТЭК России в рамках нормативно-правового и методического обеспечения в области технической защиты информационных ресурсов утвержден ряд документов [10 - 19].

Список литературы:

1. Стратегия национальной безопасности Российской Федерации (утв. указом Президента Российской Федерации от 2 июля 2021 г. № 400).
2. Доктрина информационной безопасности Российской Федерации (утв. указом Президента Российской Федерации от 5 декабря 2016 г. № 646).
3. О дополнительных мерах по обеспечению информационной безопасности Российской Федерации (утв. указом Президента Российской Федерации от 1 мая 2022 г. № 250).
4. Постановление Правительства РФ от 15 июля 2022 г. № 1272 «Об утверждении типового положения о заместителе руководителя органа (организации), ответственном за обеспечение ИБ в органе (организации), и типового положения о структурном подразделении в органе (организации), обеспечивающем информационную безопасность органа (организации)».
5. В Совбезе раскрыли детали новой доктрины информационной безопасности. - [Электронный ресурс]. URL: https://www.rbc.ru/politics/28/01/2026/6979d3ef9a79476325e0fcb9.
6. Приказ ФСТЭК России от 11 апреля 2025 г. № 117 «Об утверждении Требований о защите информации, содержащейся в государственных информационных системах, иных информационных системах государственных органов, государственных унитарных предприятий, государственных учреждений».
7. Приказ ФСТЭК России от 11 февраля 2013 г. № 17 «Об утверждении Требований о защите информации, не составляющей государственную тайну, содержащейся в государственных информационных системах».
8. Методический документ. Мероприятия и меры по защите информации, содержащейся в информационных системах. - [Электронный ресурс]. URL: https://www.fstec.ru.
9. Информационное сообщение ФСТЭК России от 12.03.2026 № 240/22/1492 «О разъяснении положений Требований о защите информации, содержащейся в государственных информационных системах, иных информационных системах государственных органов, государственных унитарных предприятий, государственных учреждений, утвержденных приказом ФСТЭК России от 11 апреля 2025 г. № 117». - [Электронный ресурс]. URL: https://fstec.ru/dokumenty/vse-dokumenty/informatsionnye-i-analiticheskie-materialy/informatsionnoe-soobshchenie-fstek-rossii-ot-12-marta-2026-g-n-240-22-1492).
10. Приказ ФСТЭК России от 4 июля 2022 г. № 118 «Об утверждении Требований по безопасности информации к средствам контейнеризации».
11. Приказ ФСТЭК России от 18 апреля 2022 г. № 68 «О внесении изменений в Требования по безопасности информации, устанавливающие уровни доверия к средствам технической защиты информации и средствам обеспечения безопасности информационных технологий, утвержденные приказом ФСТЭК России от 2 июня 2020 г. № 76».
12. Приказ ФСТЭК России от 27 октября 2022 г. № 187 «Об утверждении Требований по безопасности информации к средствам виртуализации».
13. Приказ ФСТЭК России от 01 декабря 2023 г. № 240 "Об утверждении Порядка проведения сертификации процессов безопасной разработки программного обеспечения СЗИ.
14. Методика тестирования обновлений безопасности программных, программно-аппаратных средств. Утверждена ФСТЭК России 28 октября 2022 г.
15. Методика оценки уровня критичности уязвимостей программных, программно-аппаратных средств. Утверждена ФСТЭК России 30 июня 2025 г.
16. Руководство по организации процесса управления уязвимостями в органе (организации). Утверждено ФСТЭК России 17 мая 2023 г.

17. Методика оценки уровня критичности уязвимостей программных, программно-аппаратных средств. Утверждена ФСТЭК России 30 июня 2025 г.
18. Методика оценки показателя состояния технической защиты информации в информационных системах и обеспечения безопасности ЗО КИИ РФ. Утверждена ФСТЭК России 11 ноября 2025 г.
19. Методика анализа защищенности информационных систем. Утверждена ФСТЭК России 25 ноября 2025 г.

Содержание