УДК 004.056.53

Мурзагареев Айдар Динарович, студент 4 курса бакалавриата, направление подготовки 10.03.01 «Информационная безопасность» Кафедры вычислительной техники и защиты информации, ФГБОУ ВО «Уфимский университет науки и технологий», г. Уфа

Кладов Виталий Евгеньевич, канд. техн. наук, доцент кафедры вычислительной техники и защиты информации, ФГБОУ ВО «Уфимский университет науки и технологий», Россия, г. Уфа

ИНТЕЛЛЕКТУАЛЬНАЯ ПЛАТФОРМА АВТОМАТИЗИРОВАННОГО АНАЛИЗА ВРЕДОНОСНОГО ПРОГРАММНОГО ОБЕСПЕЧЕНИЯ: ИНТЕГРАЦИЯ ФРЕЙМВОРКА KARTON, ML-КЛАССИФИКАТОРА И LLM-АССИСТЕНТОВ ЧЕРЕЗ ПРОТОКОЛ MCP

Аннотация

В статье описана архитектура платформы, объединяющей открытый фреймворк оркестрации Karton (CERT.PL, 2020), собственный модуль машинного обучения для классификации семейств вредоносного программного обеспечения и LLM-ассистента, подключаемого через протокол Model Context Protocol (Anthropic, ноябрь 2024). Актуальность работы подтверждена статистикой: AV-TEST в 2025 году фиксировал около 450 тысяч новых вредоносных файлов в сутки, «Лаборатория Касперского» за период ноябрь 2024 – октябрь 2025 называет цифру около 500 тысяч и рост 7% год к году. Сигнатурное детектирование при таких объемах перестает справляться, что подтверждено отчетами SonicWall и CrowdStrike за 2025 год. Предложенная архитектура распараллеливает статический и динамический анализ, обогащает результаты ML-предсказаниями на признаках EMBER2018 и EMBER2024 и допускает интерактивный диалог аналитика с языковой

моделью через MCP-инструменты – по аналогии с открытыми проектами ida-pro-mcp и GhidraMCP, работающими с весны 2025 года. Сделано сравнение трех режимов работы аналитика: ручного реверса, классической автоматизации и предложенного гибридного подхода. Обозначены ограничения метода, включая риски prompt injection в LLM-интеграции, и направления дальнейшей работы.

Annotation

The article describes the architecture of a platform that combines the open-source orchestration framework Karton (CERT.PL, 2020), a custom machine learning module for malware family classification, and an LLM assistant connected via the Model Context Protocol (Anthropic, November 2024). The relevance is supported by statistics: AV-TEST registered approximately 450,000 new malicious files per day in 2025, while Kaspersky reports an average of about 500,000 per day for the November 2024 – October 2025 period, a 7% year-over-year increase. Signature-based detection no longer scales at such volumes, as confirmed by SonicWall and CrowdStrike 2025 reports. The proposed architecture parallelizes static and dynamic analysis, enriches results with ML predictions on EMBER2018 and EMBER2024 features, and supports interactive analyst-to-LLM dialogue through MCP tools, analogous to the open-source projects ida-pro-mcp and GhidraMCP that have been in active use since spring 2025. A comparison of three analyst workflows is provided: manual reversing, classical automation, and the proposed hybrid approach. Limitations of the method, including prompt injection risks in LLM integration, and directions for future work are outlined.

Ключевые слова: *вредоносное программное обеспечение, реверс-инжиниринг, машинное обучение, Karton, MCP, классификация семейств, автоматизация анализа, большие языковые модели*.

Keywords: *malware, reverse engineering, machine learning, Karton, MCP, family classification, automated analysis, large language models*.

Введение

AV-TEST в 2025 году регистрировал в среднем 450 тысяч новых вредоносных и потенциально нежелательных файлов в сутки [1]. «Лаборатория Касперского» за период с ноября 2024 по октябрь 2025 называет цифру около 500 тысяч и рост 7% к предыдущему окну [2]. Источники расходятся, но порядок один – полмиллиона в день.

Ручной реверс при таких потоках не масштабируется. Аналитик уровня Senior тратит на нетривиальный семпл часы, иногда сутки. Точных публичных бенчмарков по среднему времени разбора одного образца в индустрии нет – компании эти метрики не раскрывают, и в настоящей работе они оцениваются качественно.

Сигнатурное детектирование справляется хуже, чем принято считать. SonicWall в годовом отчете 2025 фиксирует более 210 тысяч ранее невиданных вариантов ВПО за один лишь 2024 год [3]. CrowdStrike в обзоре 2025 утверждает, что 79% удачных взломов идут вообще без вредоносного файла – через украденные учетные данные и living-off-the-land [4]. Цифра спорная: разные вендоры считают «атаку» по-разному, но направление тренда устойчиво.

Параллельно сформировался новый инструмент аналитика – большие языковые модели. Pordanesh и Tan в работе 2024 года проверили GPT-4 на бинарном реверсе: модель уверенно интерпретирует декомпилированный код общего назначения, но проседает на детальном security-анализе [5]. Patsakis, Casino и Lykousas в том же 2024 году протестировали четыре LLM на реальных скриптах кампании Emotet – модели частично, но небезошибочно справились с деобфускацией [6]. Обзор Jelodar и соавторов 2025 года на 127 публикациях подтверждает общую картину: LLM ускоряют рутину, но не заменяют аналитика [7].

Anthropic 25 ноября 2024 года опубликовала Model Context Protocol (MCP) – открытый стандарт подключения LLM к внешним инструментам [8]. За первый год существования протокол собрал свыше 10 тысяч активных серверов и был передан в декабре 2025 года Linux Foundation [8]. Для реверсера это значит, что IDA Pro, Ghidra и пользовательские пайплайны становятся вызываемыми из чата напрямую.

Цель работы – описать архитектуру платформы из трех слоев: оркестрация конвейера через Karton, ML-классификация семейств ВПО и LLM-ассистент через MCP. Задачи – разобрать существующие классы решений, обосновать выбор Karton, описать ML-модуль и сценарий MCP-интеграции, сравнить варианты по эксплуатационным характеристикам.

Существующие подходы и место Karton

Поле автоматизированного анализа ВПО неоднородно. Имеющиеся решения распадаются на четыре класса – по тому, какую часть работы они автоматизируют.

Песочницы – Cuckoo Sandbox, Joe Sandbox, Any.Run – исполняют семпл в изолированной среде и собирают поведенческие индикаторы: системные вызовы, сетевой трафик, файловые операции. Преимущество – обнаружение по поведению, а не по сигнатуре. Слабость известна: современные образцы используют anti-VM и anti-debug, обнаруживают среду и не активируют полезную нагрузку.

Платформы агрегации индикаторов – MWDB Core и MISP – хранят и распространяют метаданные: хеши, YARA-правила, артефакты. Анализ они не выполняют; их роль – коллективная защита.

Коммерческие облачные платформы – VirusTotal, Hybrid Analysis, Intezer Analyze – дают глубокий анализ, но загруженные образцы перестают быть конфиденциальными. Для корпоративных и государственных инцидентов это нередко неприемлемо.

Karton занимает в этой картине особое место. CERT.PL открыла исходный код фреймворка в декабре 2020 года [9]. Сам Karton ничего не анализирует – это распределенный оркестратор задач: брокер Redis, объектное хранилище MinIO, набор Python-микросервисов, обменивающихся типизированными сообщениями. Анализ выполняют независимые сервисы, которые администратор подключает или пишет самостоятельно. Эта модульность и стала причиной выбора – пользовательский ML-сервис и MCP-мост подключаются без модификации ядра.

Архитектура предлагаемой платформы

Конвейер выстроен в три фазы: первичная классификация, параллельный анализ, агрегация и интерактивная экспертиза. Структурная схема приведена на рисунке 1.

На вход подается семпл. Сервис karton-classifier определяет тип образца – PE, ELF, документ, скрипт – и порождает три параллельные задачи.

Первая ветка – статический анализ: разбор PE-заголовков, секций, импортов, строк, оценка энтропии. Вторая – запуск в песочнице и снятие поведенческих индикаторов. Третья – собственный ML-сервис, классифицирующий семейство ВПО по статическим признакам.

Все три результата стекаются в MWDB Core и формируют обогащенный профиль образца. Параллельно к платформе подключен MCP-сервер: он экспонирует ключевые функции – получение профиля, запрос ML-предсказания, прогон YARA-правила, выдачу фрагмента дизассемблера – в виде инструментов для LLM-клиента. Разделение ответственности явное: конвейер берет рутину, LLM обслуживает интерактивную экспертизу.

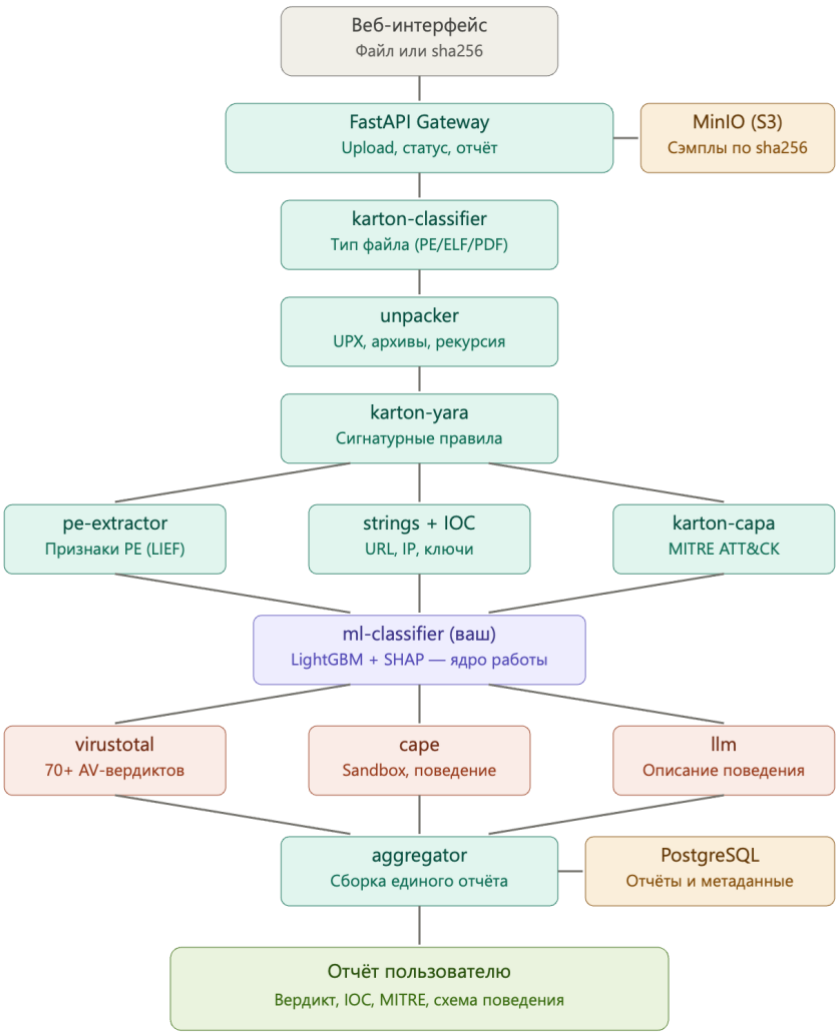

Рисунок 1. Структурная схема интеллектуальной платформы автоматизированного анализа ВПО

Модуль машинного обучения

Задача ML-модуля – мультиклассовая классификация семейств ВПО по статическим признакам PE-файлов.

В качестве обучающего набора используется EMBER. Версия 2018 (Anderson и Roth) – 1,1 миллиона образцов и 2381 признак, ставшая фактическим стандартом для статического анализа [10]. В 2025 году Joyce и соавторы представили EMBER2024 – уже 3,2 миллиона файлов шести

форматов, отдельный «challenge»-набор из образцов, обходивших антивирусы, и поддержка задачи family classification из коробки [11]. Имеет смысл сразу строиться на EMBER2024 – он шире и созданее.

Базовая модель – градиентный бустинг LightGBM. Сравнительные эксперименты на EMBER показывают точность LightGBM и XGBoost выше 96% для бинарной классификации и устойчивое преимущество над KNN и TabNet [11]. Время предсказания – единицы миллисекунд на образец, что критично для высоконагруженного конвейера. Глубокие модели типа MalConv в MVP избыточны: прирост качества не оправдывает усложнение инференса. К ним разумно возвращаться после набора статистики на реальной нагрузке.

В платформу модель встраивается как Karton-сервис, наследующий karton.core.Karton. Сервис слушает задачи с тегом kind: runnable-pe, извлекает признаки через lief и pefile, прогоняет их через сохраненную модель и публикует сообщение feed:malware-family со структурой {family, confidence, top_3}. Переобучение и обновление модели не требуют остановки остальных сервисов.

Интеграция LLM-ассистента через MCP

MCP – открытый стандарт связи LLM с внешними системами, опубликованный Anthropic 25 ноября 2024 года [8]. До MCP каждая пара «модель – инструмент» требовала собственного коннектора. После – одна реализация на каждом конце.

MCP-сервер платформы экспонирует четыре инструмента: get_sample_info(sha256) – профиль из MWDB, get_ml_prediction(sha256) – результат классификации, run_yara(rule, sha256) – прогон правила, get_disassembly(sha256, address, length) – фрагмент кода. Список расширяется без изменений на клиентской стороне.

В реверс-сообществе уже работают два открытых проекта, подтверждающих жизнеспособность подхода. Проект ida-pro-mcp Дункана Огилви (mrexodia) подключает IDA Pro к LLM через MCP и позволяет

переименовывать функции и переменные, ставить комментарии и навигироваться по декомпиляции командами на естественном языке [12]. Проект GhidraMCP Лори Уайред решает ту же задачу для Ghidra [13]. Оба активно развиваются с весны 2025 года; ida-pro-mcp включен в официальный репозиторий плагинов Hex-Rays.

Тот же опыт показывает и пределы метода. В документации ida-pro-mcp авторы прямо предупреждают: LLM плохо считает в уме и галлюцинирует на обфусцированном коде – преобразования между основаниями систем счисления приходится делегировать отдельному инструменту int_convert, а упаковщики и библиотечные функции снимать до запроса к модели [12]. Рабочий сценарий выглядит так: аналитик пишет в чате «разбери функцию 0x401A20 в семпле с хешем X», модель последовательно вызывает get_sample_info и get_disassembly, формирует интерпретацию с указанием подозрительных API-вызовов, аналитик принимает финальное решение. Атрибуция и выводы остаются за человеком – LLM работает интерфейсом, не оракулом.

Сравнительный анализ

Сравнение трех режимов работы при одной и той же задаче – классификация и описание поведения нового неизвестного образца ВПО – сведено в таблицу 1.

Таблица 1. Сравнение режимов работы аналитика

Критерий	**Ручной реверс**	**Классическая автоматизация**	**Гибридный подход (Karton + ML + LLM/MCP)**
Среднее время на образец	часы – сутки	5–15 минут	1–3 минуты
Требуемая квалификация	Senior RE	Middle / Junior+	Junior на типовых случаях
Покрытие обфусцированных образцов	высокое	среднее	среднее

Горизонтальная масштабируемость	низкая	высокая	высокая
Объяснимость результата	высокая	низкая	высокая (LLM-комментарии)
Конфиденциальность образца	высокая	высокая (on-prem)	средняя (запросы к LLM)

Цифры в столбцах «время» – ориентировочные. Открытых индустриальных бенчмарков по среднему времени разбора одного семпла нет, оценки опираются на публикации [5, 7] и стандартные ожидания SOC-практики. Точные сравнения возможны только на собственной выборке и собственном секундомере.

Тенденция при этом видна и без точных цифр: гибридный режим сокращает время рутинного разбора и опускает порог квалификации линейного аналитика. Сложные нетиповые случаи остаются за Middle-Senior инженерами – их этот режим не вытесняет.

Заключение

Описана архитектура платформы из трех слоев: Karton как оркестратор конвейера, собственный ML-сервис на признаках EMBER для классификации семейств ВПО и MCP-мост, экспонирующий инструменты платформы LLM-ассистенту.

Жизнеспособность MCP-интеграции в реверсе уже подтверждена практикой – ida-pro-mcp и GhidraMCP работают с весны 2025 года и используются в реальной аналитической работе [12, 13]. Перенос той же модели на платформенный уровень – логичный следующий шаг.

Дальнейшие направления – расширение набора форматов (ELF, документы Office, скрипты), переход с EMBER2018 на EMBER2024 [11] для оценки на «challenge»-наборе антивирус-устойчивых образцов и эксперименты с глубокими моделями на обфусцированных семплах. Отдельный блок – безопасность самой LLM-интеграции. В апреле 2025 года

исследователи опубликовали анализ MCP с выводом о множественных нерешенных проблемах: prompt injection, права инструментов, допускающие комбинаторное извлечение данных, и подменные инструменты-двойники [8]. Содержимое исследуемых семплов в нашем сценарии – не доверенный текст, и эти риски нужно закладывать в дизайн.

Список литературы

1. AV-TEST Institute. Malware Statistics & Trends Report. URL: https://www.av-test.org/en/statistics/malware/ (дата обращения: 05.05.2026).

2. Kaspersky Security Bulletin 2025. Statistics // Securelist. – 02.12.2025. URL: https://securelist.com/kaspersky-security-bulletin-2025-statistics/118189/ (дата обращения: 05.05.2026).

3. SonicWall 2025 Annual Cyber Threat Report. URL: https://www.sonicwall.com/threat-report (дата обращения: 05.05.2026).

4. CrowdStrike Global Threat Report 2025. URL: https://www.crowdstrike.com/global-threat-report/ (дата обращения: 05.05.2026).

5. Pordanesh S., Tan B. Exploring the Efficacy of Large Language Models (GPT-4) in Binary Reverse Engineering // arXiv preprint arXiv:2406.06637. – 2024. URL: https://arxiv.org/abs/2406.06637.

6. Patsakis C., Casino F., Lykousas N. Assessing LLMs in Malicious Code Deobfuscation of Real-world Malware Campaigns // arXiv preprint arXiv:2404.19715. – 2024. URL: https://arxiv.org/abs/2404.19715.

7. Jelodar H., Bai S., Hamedi P. et al. Large Language Model (LLM) for Software Security: Code Analysis, Malware Analysis, Reverse Engineering // arXiv preprint arXiv:2504.07137. – 2025. URL: https://arxiv.org/abs/2504.07137.

8. Anthropic. Introducing the Model Context Protocol. – 25.11.2024. URL: https://www.anthropic.com/news/model-context-protocol (дата обращения: 05.05.2026).

9. CERT Polska. Karton: Distributed malware processing framework. – Открыт в декабре 2020. URL: https://github.com/CERT-Polska/karton (дата обращения: 05.05.2026).

10. Anderson H.S., Roth P. EMBER: An Open Dataset for Training Static PE Malware Machine Learning Models // arXiv preprint arXiv:1804.04637. – 2018. URL: https://arxiv.org/abs/1804.04637.

11. Joyce R. et al. EMBER2024 – A Benchmark Dataset for Holistic Evaluation of Malware Classifiers // Proceedings of the 31st ACM SIGKDD Conference on

Knowledge Discovery and Data Mining. – 2025. URL: https://dl.acm.org/doi/10.1145/3711896.3737431.

12. Ogilvie D. (mrexodia). ida-pro-mcp: AI-powered reverse engineering assistant. – 2025. URL: https://github.com/mrexodia/ida-pro-mcp (дата обращения: 05.05.2026).

13. Wired L. GhidraMCP: MCP Server for Ghidra. – 2025. URL: https://github.com/LaurieWired/GhidraMCP (дата обращения: 05.05.2026).

References

1. AV-TEST Institute. Malware Statistics & Trends Report. URL: https://www.av-test.org/en/statistics/malware/ (accessed 05.05.2026).

2. Kaspersky Security Bulletin 2025. Statistics. Securelist, December 2, 2025. URL: https://securelist.com/kaspersky-security-bulletin-2025-statistics/118189/ (accessed 05.05.2026).

3. SonicWall 2025 Annual Cyber Threat Report. URL: https://www.sonicwall.com/threat-report (accessed 05.05.2026).

4. CrowdStrike Global Threat Report 2025. URL: https://www.crowdstrike.com/global-threat-report/ (accessed 05.05.2026).

5. Pordanesh S., Tan B. Exploring the Efficacy of Large Language Models (GPT-4) in Binary Reverse Engineering. arXiv preprint arXiv:2406.06637, 2024. URL: https://arxiv.org/abs/2406.06637.

6. Patsakis C., Casino F., Lykousas N. Assessing LLMs in Malicious Code Deobfuscation of Real-world Malware Campaigns. arXiv preprint arXiv:2404.19715, 2024. URL: https://arxiv.org/abs/2404.19715.

7. Jelodar H., Bai S., Hamedi P. et al. Large Language Model (LLM) for Software Security: Code Analysis, Malware Analysis, Reverse Engineering. arXiv preprint arXiv:2504.07137, 2025. URL: https://arxiv.org/abs/2504.07137.

8. Anthropic. Introducing the Model Context Protocol. November 25, 2024. URL: https://www.anthropic.com/news/model-context-protocol (accessed 05.05.2026).

9. CERT Polska. Karton: Distributed malware processing framework. Open-sourced December 2020. URL: https://github.com/CERT-Polska/karton (accessed 05.05.2026).
10. Anderson H.S., Roth P. EMBER: An Open Dataset for Training Static PE Malware Machine Learning Models. arXiv preprint arXiv:1804.04637, 2018. URL: https://arxiv.org/abs/1804.04637.
11. Joyce R. et al. EMBER2024 – A Benchmark Dataset for Holistic Evaluation of Malware Classifiers. Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2025. URL: https://dl.acm.org/doi/10.1145/3711896.3737431.
12. Ogilvie D. (mrexodia). ida-pro-mcp: AI-powered reverse engineering assistant. 2025. URL: https://github.com/mrexodia/ida-pro-mcp (accessed 05.05.2026).
13. Wired L. GhidraMCP: MCP Server for Ghidra. 2025. URL: https://github.com/LaurieWired/GhidraMCP (accessed 05.05.2026).