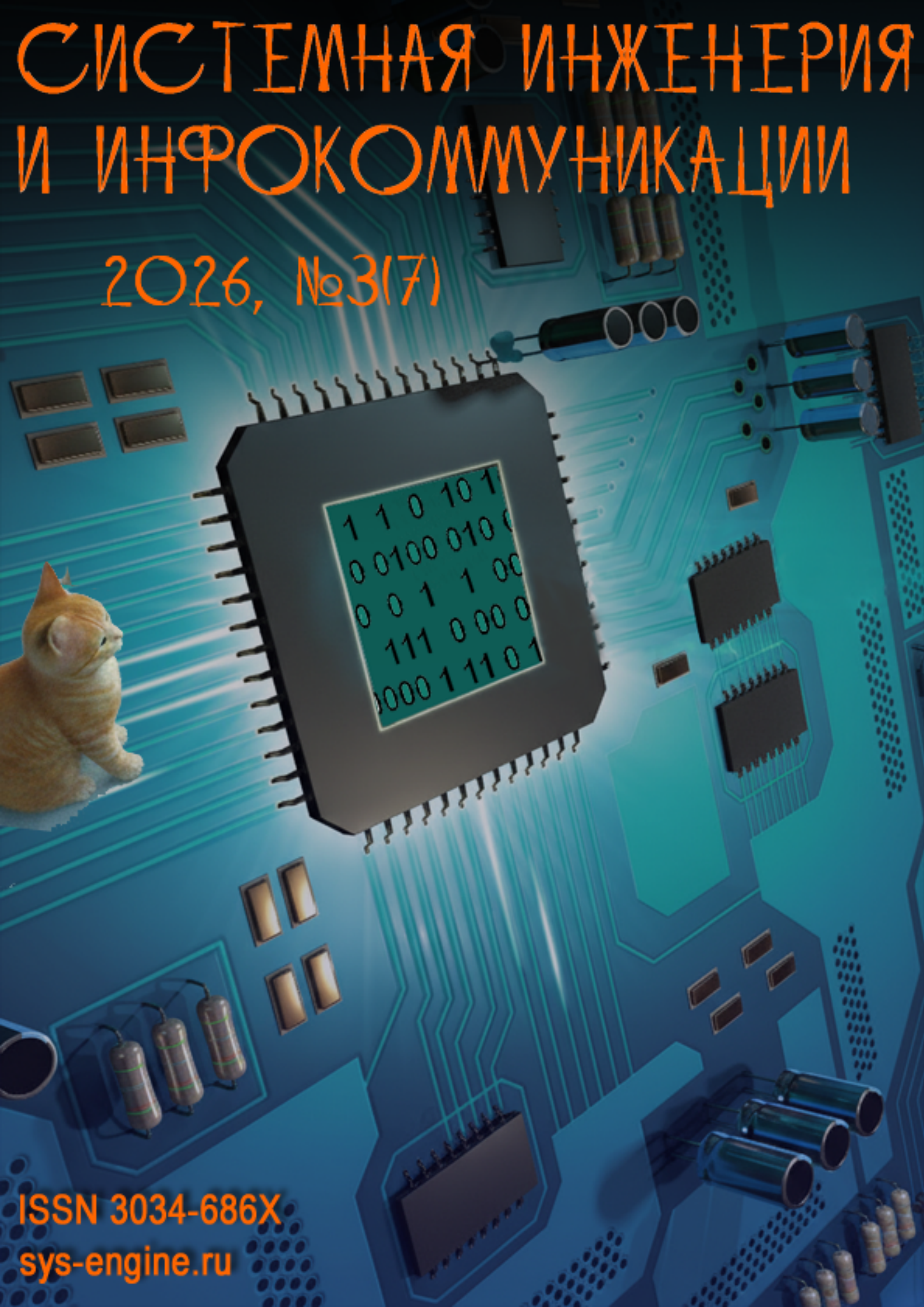
СИСТЕМНАЯ ИНЖЕНЕРИЯ И ИНФОКОММУНИКАЦИИ

2026, №3(7)



ISSN 3034-686X
sys-engine.ru

Системная инженерия и инфокоммуникации

№3(7), 2026

Регистрационный номер: Эл № ФС77-88783 от 22 ноября 2024 г.
ISSN 3034-686X
Сетевое издание размещено по адресу: sys-engine.ru
Возрастная категория: 12+

Содержание

Алгоритм предиктивного управления автономным транспортным средством в среде с препятствиями

С.А.К. Диане [1,2], *В.И. Назаркин* [2], *А.В. Бессонова* [2], *А.В. Суслова* [2]

[1]*Институт проблем управления им. В. А. Трапезникова РАН, Москва, Россия*
[2]*МИРЭА – Российский технологический университет, Москва, Россия*

Аннотация **– Ключевыми задачами при эксплуатации автономного транспортного средства, как известно, являются следование намеченному маршруту и избегание столкновений с препятствиями. В настоящей работе предложен алгоритм управления, основанный на прогнозирующей имитационной модели беспилотной платформы, который позволяет реализовать такое предиктивное поведение. В качестве методологической основы использован метод Эйлера для численного решения дифференциальных уравнений кинематики транспортного средства, метод динамического окна для оценки и выбора возможных траекторий транспортного средства при различных управляющих уставках, а также редуцированный метод нечетко-логического вывода при выборе целевых точек маршрута. Эксперименты, проведенные в виртуальной среде, подтверждают эффективность подхода на тестовых сценариях движения.**

Ключевые слова **– предиктивное управление, среда с препятствиями, метод динамического окна, автономный транспорт, нечеткая логика.**

ВВЕДЕНИЕ

Задача организации пассажирских и грузовых перевозок является критически важной при обеспечении жизнедеятельности крупных городов, промышленных предприятий, организаций розничной торговли. По последним оценкам транспортная сфера формирует около 6% валового внутреннего продукта России [1].

Наряду с созданием и внедрением в повседневную жизнь беспилотных транспортных средств (БТС) для повышения оперативности и безопасности городского дорожного движения активное внимание к разработке автономных транспортных платформ (АТП) проявляется со стороны владельцев крупных складских предприятий.

В обоих случаях существенной научно-технической проблемой в решении транспортно-логистических задач является сложность организации безопасного маневрирования транспортных средств в условиях нелинейности и неголономности их кинематики, динамичности и неполнозаданности внешней среды.

Существующие подходы к решению обозначенной проблемы опираются на принципы ситуационного управления [2, 3], графовые методы планирования движений [4], методы мультимодальной фильтрации цифровых сигналов [5], технологии навигации [6], методы адаптивного регулирования [7], технологии интеллектуальной обработки информации [8], а также алгоритмы предиктивного анализа данных [9, 10].

Целью настоящей работы является разработка математической формализации для моделирования задачи целенаправленного движения АТП в среде с препятствиями и совместимого с ней алгоритма управления, учитывающего параметры платформы и внешней среды.

В основе предлагаемого подхода лежит метод динамического окна [11, 12] и аналитические принципы оценки пересечения геометрических фигур [13]. При этом преимуществами разработанного алгоритма являются применение редуцированного нечетко-логического вывода для оценки значимости рассогласований до ближайших отрезков маршрута и учет полигональной формы подвижных объектов на пути следования АТП.

I. ПОСТАНОВКА ЗАДАЧИ

В целях обеспечения достоверности и повторяемости результатов исследования опишем математическую модель АТП и среды ее перемещения, допускающую применение метода Эйлера для численного решения дифференциальных уравнений кинематики.

Рассмотрим четырехколесную АТП с рулевой рейкой в передней части. Выберем за основу широко известную велосипедную кинематику [14]:

$$\begin{aligned} x_{t+1} &= x_t + v\ cos(\alpha)\Delta t, \\ y_{t+1} &= y_t + v\ sin(\alpha)\Delta t, \\ \alpha_{t+1} &= \alpha_t + \frac{v\ tg(\theta)}{L}\Delta t, \end{aligned} \tag{1}$$

где x_t, y_t – координаты центра платформы на плоскости, α_t – угол поворота платформы, v – скорость движения платформы, θ – угол поворота рулевых колес, Δt – шаг расчета.

Зададим маршрут G движения АТП, состоящий из K опорных точек, и фактическую траекторию робота H, пополняемую с течением времени t, в виде следующих множеств:

$$G = \{g_1, \ldots, g_K\} \in \mathbb{R}^{2K}, \quad (2)$$

$$H = \{h_1, \ldots, h_t\} \in \mathbb{R}^{2t}. \quad (3)$$

Корпус робота P и геометрию препятствий Q внешней среды определим, соответственно, в формате многоугольников

$$P = \{p_1, \ldots, p_M\} \in \mathbb{R}^{2M}, \quad (4)$$

$$Q = \{q_1, \ldots, q_N\} \in \mathbb{R}^{2N}, \quad (5)$$

вершины которых перечислены в порядке обхода против часовой стрелки.

Учтем также возможность задания внешней трансформации препятствия (5) с указанием координат центральной точки и угла поворота многоугольника:

$$\Theta_Q = \{x_Q, y_Q, \beta\}. \quad (6)$$

Дополнительно с каждым препятствием (5) в рабочей среде АТП ассоциируем параметры движения – его линейную и угловую скорости:

$$\Phi_Q = \{v_x, v_y, \omega\}. \quad (7)$$

На Рис. 1 показана схема движения платформы в среде с препятствиями. Наряду с основными геометрическими элементами указаны ошибки e_t рассогласования АТП с маршрутом и точка p_c потенциального столкновения с препятствием.

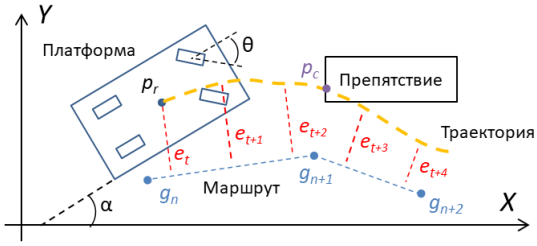


Рис. 1. Схема целенаправленного движения робота в среде с препятствиями

Требуется решить задачу: обеспечить движение АТП по траектории H близкой к маршруту G при одновременном уклонении от одного или нескольких подвижных препятствий $Q^+ = \{Q_1, \ldots, Q_J\}$.

Отметим, что базовый закон управления транспортной платформой может быть основан на пропорциональном регулировании угла поворота рулевых колес:

$$e_\alpha = (atan2(y_g - y_r, x_g - x_r) - \alpha_t) \rightarrow [-\pi; \pi), \quad (8)$$

$$u_\theta = k_p e_\alpha + k_d \dot{e}_\alpha, \quad (9)$$

$$\Delta\theta = u_\theta \Delta t, \quad (10)$$

где e_α – угловое рассогласование, (x_g, y_g) – координаты целевой точки, (x_r, y_r) – координаты робота, u_θ – управляющая уставка, k_p – коэффициент пропорционального регулирования, k_d – коэффициент дифференциального регулирования, $\Delta\theta$ – приращение угла за такт моделирования, $\rightarrow$ – операция приведения угла к стандартному диапазону на тригонометрической окружности.

Однако применение данного базового закона управления дает высокую ошибку следования траектории и не учитывает наличие подвижных препятствий – следовательно, требуется более эффективный метод.

II. ПРЕДИКТИВНЫЙ АЛГОРИТМ ПЛАНИРОВАНИЯ ТРАЕКТОРИИ

Улучшенный предиктивный закон управления предлагается строить на предположении о возможности рекуррентного прогноза состояния АТП $s_t = \{x_t, y_t, \alpha_t\}$ на несколько шагов времени вперед:

$$s_{t+1} = f_{robot}(s_t, u_t), \quad (11)$$

где f_{robot} реализуется на основе (1), а также на основе аналогичного прогноза состояний (трансформаций) препятствий:

$$\Theta_{Q,t+1} = f_{obstacle}(\Theta_{Q,t}, \Phi_{Q,t}), \quad (12)$$

где $f_{obstacle}$ реализуется на основе законов равномерного движения с учетом параметров (6), (7).

Входом алгоритма являются начальное состояние робота s_{t0}, целевой маршрут движения G, геометрия ближайших препятствий внешней среды Q^+, а также их параметры движения Θ_Q, Φ_Q.

На выходе алгоритма в каждый дискретный момент времени определяется серия управляющих воздействий:

$$U_\theta = \arg\min_{U_\theta} \sum_{t=t_0+1}^{t_0+N_h} e_t\left(s_t(s_{t_0}, u_{t_0}, \ldots, u_{t-1}), G, P, Q^+\right), \quad (13)$$

$$U_\theta = \{u_{t_0}, \ldots, u_{t_0+N_h-1}\} \in \mathbb{R}^{2N_h}, \quad (14)$$

где u_t – уставки по скорости движения и углу руления, N_h – горизонт прогноза, G – маршрут движения, P – геометрия корпуса платформы, Q^+ – геометрия ближайших к АТП препятствий внешней среды.

В рамках основных этапов алгоритма проводится решение оптимизационной задачи (13) с численным моделированием робота по формуле (11) и динамических препятствий по формуле (12).

Далее выполняется управление u_{t_0}, соответствующее первому элементу множества (14). Последующие же уставки отбрасываются в ожидании перепланирования движений АТП на очередном шаге моделирования.

Уточненное выражение для ошибки в каждой точке $h_t = p_r(t)$ движения робота имеет следующий вид:

$$e_t = k_1 \lambda(h_t, g_{n-1}, g_n) + k_2 \lambda(h_t, g_n, g_{n+1}) + \rho_G + \rho_Q, \quad (15)$$

где g_n – опорная точка маршрута, λ – геометрическая функция вычисления расстояния от точки до отрезка,

ρ_G – некоторый штраф за пропуск опорной точки маршрута, ρ_Q – некоторый штраф за столкновение с препятствием, k_1 – коэффициент стремления к текущей опорной точке, k_2 – коэффициент стремления к следующей опорной точке.

Расчет коэффициентов k_1, k_2 проводится на основе упрощенной реализации нечетко-логических правил:

«Если робот ближе к точке g_n чем к точке g_{n+1}, то k_1 – высокий, k_2 – низкий»;

«Если робот ближе к точке g_{n+1} чем к точке g_n, то k_1 – низкий, k_2 – высокий».

Алгебраическая реализация данных правил выполнена следующим образом:

$$k_1 = \frac{\|h_t - g_{n+1}\|^2}{\|h_t - g_n\|^2 + \|h_t - g_{n+1}\|^2}, \quad (16)$$

$$k_2 = 1 - k_1. \quad (17)$$

Событие пропуска точки маршрута отслеживается в случае увеличения дистанции от АТП до этой точки, но недостижении порогового расстояния ε_G на недавнем отрезке времени:

$$\begin{cases} \min\limits_{t^* \in [t_{prev}, t]} \|h_{t^*} - g_n\| > \varepsilon_G, \\ \dfrac{d}{dt}\|h_t - g_n\| > 0, \end{cases} \quad (18)$$

где t_{prev} – время прохождения прошлого отрезка маршрута или начала моделирования.

Критерием опасности сближения объектов (АТП и препятствия) считается уменьшение кратчайшего расстояния D между точками их многоугольников ниже критической отметки.

Оценка взаимного расположения многоугольников опирается на расчет функции для проверки круговой упорядоченности трех произвольных точек a, b и c:

$$\zeta(p_a, p_b, p_c) = (y_c - y_a)(x_b - x_a) > (y_b - y_a)(x_c - x_a), \quad (19)$$

которая принимает значение 1, если точки идут против часовой стрелки и 0, иначе.

Рассмотрим три типовых случая взаимного расположения многоугольников (Рис. 2): (1) вхождение нескольких точек одного многоугольника в другой, (2) наложение двух многоугольников без вхождений точек и (3) соседство многоугольников.

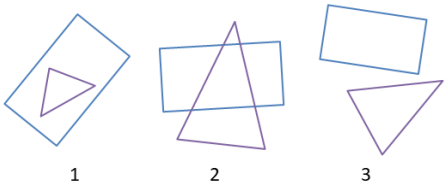


Рис. 2. Случаи взаимного расположения двух многоугольников: 1 – вхождение, 2 - наложение, 3 – соседство

Для 1-го случая примем D=0. Состоятельность 1-го случая определяется фактом ξ принадлежности некоторой точки p^* многоугольнику Q, что в свою очередь зависит от четности числа δ фактов сонаправленности ориентированных углов обхода этой точки с общим направлением обхода многоугольника:

$$\xi(p^*, Q) = \begin{cases} 1, \textit{ если } \sum\limits_{i=1}^{N+1} \delta(p^*, Q, i) \in \{2n\},\ n \in \mathbb{N}, \\ 0, \textit{ иначе}, \end{cases} \quad (20)$$

$$\delta(p^*, Q, i) = \begin{cases} \zeta(p^*, q_i, q_{i+1}), \textit{ если} \\ \min(y_{q_i}, y_{q_{i+1}}) \le y^* < \max(y_{q_i}, y_{q_{i+1}}), \\ 0, \textit{ иначе}, \end{cases} \quad (21)$$

где i = 1, ..., N, q_{N+1} = q_0; причем углы обхода образуются лучами p^*q_i, и p^*q_{i+1}, проходящими лишь через те вершины многоугольника Q, которые проекционно связаны с точкой p^* по оси ординат (для сокращения числа вычислений).

Для 2-го случая также примем D=0. Состоятельность 2-го случая определяется фактом пересечения χ хотя бы одного отрезка (p_a, p_b), принадлежащего первому многоугольнику и отрезка (p_c, p_d) в составе второго многоугольника:

$$\exists (p_a, p_b) \in P, (p_c, p_d) \in Q:$$

$$\chi(p_a, p_b, p_c, p_d) = \zeta(p_a, p_c, p_d) \ne \zeta(p_b, p_c, p_d) \wedge \quad (22)$$

$$\wedge\, \zeta(p_a, p_b, p_c) \ne \zeta(p_a, p_b, p_d) = 1.$$

Для 3-го случая D определяется по формуле ортогонального расстояния от точки p^* до отрезка (p_a, p_b):

$$D(p^*, p_a, p_b) = \begin{cases} |x^* - x_a|, \textit{ если } x_a = x_b, \\ \dfrac{|k(x^* - x_a) - (y^* - y_a)|}{\sqrt{k^2 + 1}}, k = \dfrac{y_b - y_a}{x_b - x_a}, \textit{ иначе.} \end{cases} \quad (23)$$

Из всевозможных сочетаний точек первого и второго многоугольников следует выбрать то, которое дает минимальное расстояние:

$$D_{\min} = \min\Big(\min_{\substack{p \in P, q_a \in Q, \\ q_b \in Q}} D(p, q_a, q_b), \min_{\substack{q \in Q, p_c \in P, \\ p_d \in P}} D(q, p_c, p_d)\Big). \quad (24)$$

Полученная оценка расстояния между многоугольниками используется в формуле (15) для расчета штрафа ρ_Q и способствует выбору безопасной траектории движения АТП.

III. РЕЗУЛЬТАТЫ ЭКСПЕРИМЕНТОВ

Методика проведения экспериментальных исследований заключалась в программной реализации предложенного алгоритмического обеспечения; в сравнении базового режима движения (на основе траекторных П-регулятора и ПД-регулятора) с усовершенствованным предиктивным алгоритмом управления транспортным средством; а также в формировании выводов о применимости подхода.

Программная реализация классических и предиктивного режимов движения АТП выполнена на языке Python 3 с применением библиотек numpy, pygame.

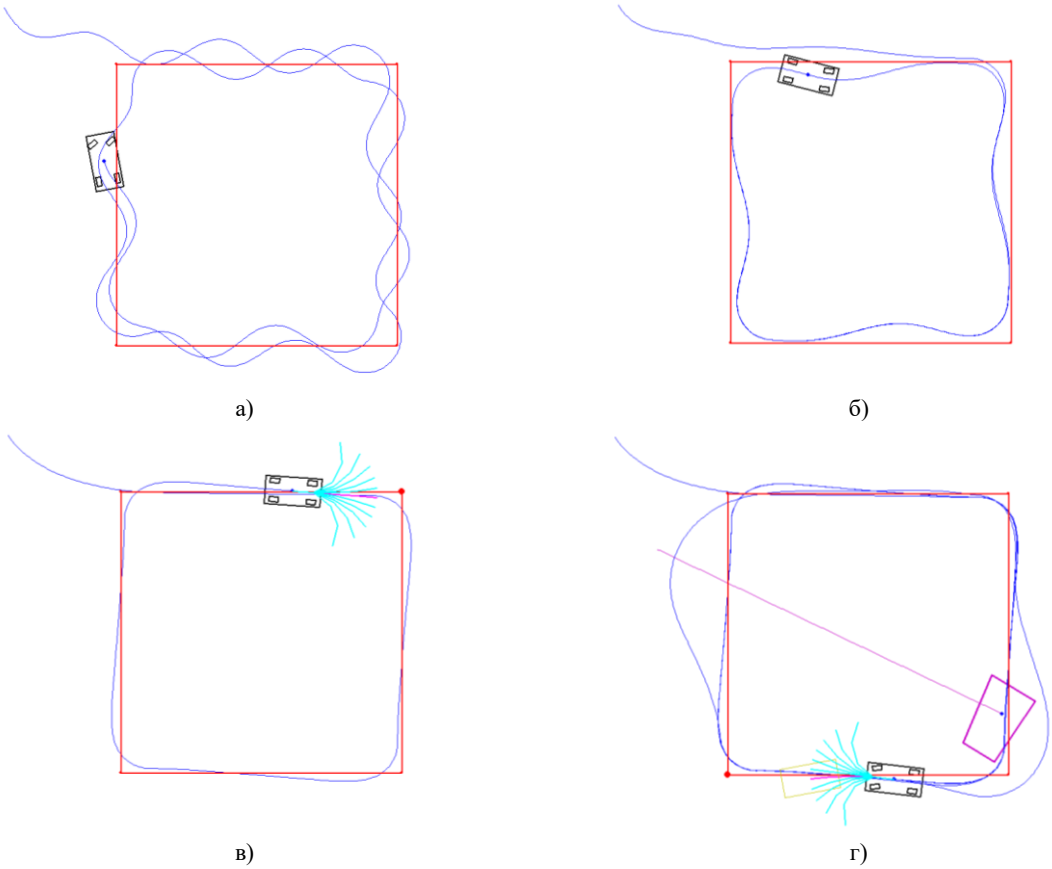


Рис. 3. Траектории АТП при различных режимах и условиях управления: а) траекторное П-регулирование; б) траекторное ПД-регулирование; в) предиктивное управление без препятствий; г) предиктивное управление с динамическим препятствием

Моделировалась АТП размером 4м × 2м, перемещающаяся по циклической траектории в форме квадрата размером 20м × 20м. Тестовая скорость движения АТП была выбрана равной 1 м/с.

Для режима П-регулирования была получена траектория, представленная на Рис. 3, а. Видны значительные отклонения от маршрута

Для режима ПД-регулирования была получена траектория, представленная на Рис. 3, б. Гладкость движений повысилась, а ошибка следования маршруту уменьшилась.

Для режима предиктивного управления в среде без препятствий была получена траектория, представленная на Рис. 3, в. Качество траектории существенно выше, чем при базовом подходе. Представленный на рисунке веер потенциальных маршрутов отражает специфику применения метода динамического окна, где из нескольких альтернатив по управляющим уставкам робот выбирает наилучшую.

Для режима предиктивного управления в среде с динамическим препятствием была получена траектория, представленная на Рис. 3, г. Качество траектории не уступает предыдущему эксперименту. Отклонения от маршрута обусловлены требованием безопасности движения вблизи препятствия (показано фиолетовым цветом). Бежевым цветом показано прогнозируемое положение робота согласно выбранной альтернативе движения на горизонте прогнозирования 80 шагов (4 секунды). Отметим, что для запечатленного момента времени (292 секунда моделирования) отсутствие пересечений с многоугольником препятствия обеспечивает нулевой штраф ρ_Q в ошибке (15). Равным образом нулевое значение имеет штраф ρ_G, ввиду приближения АТП к очередной позиции (красная точка на Рис. 3, г).

Таким образом, можно сформировать вывод о применимости и эффективности предложенного подхода к предиктивному управлению с многокритериальной оценкой ошибки на базе редуцированной схемы нечетко-логического вывода и геометрического подхода к вычислению расстояний между ограничивающими многоугольниками объектов.

ВЫВОДЫ И ЗАКЛЮЧЕНИЕ

В ходе работы была предложена математическая формализация подхода к предиктивному управлению автономной транспортной платформой, основанного на комплексном использовании метода Эйлера численного решения дифференциальных уравнений в части расчета велосипедной кинематики транспортного

средства, методов динамического окна и нечеткой логики, а также элементов аналитической геометрии в части оценки потенциальных траекторий движения АТП.

Проведённые экспериментальные исследования полностью подтверждают применимость данного подхода и его повышенную эффективность при отработке траекторий в формате ломаных линий, в т.ч. замкнутого типа. Предложенный алгоритм хорошо справляется с уклонением от динамических препятствий.

Дальнейшие перспективы развития исследования сопряжены с разработкой алгоритма адаптивного выбора порога прогнозирования, а также расширения постановки задачи до полномасштабной навигации АТП в пределах моделируемого фрагмента городской среды. Определенный интерес представляет комбинирование режимов ПД-регулирования и предиктивного управления.

ЛИТЕРАТУРА

[1] Транспортный комплекс входит в пятерку крупнейших российских отраслей // Аналитический центр при Правительстве Российской Федерации. URL: https://ac.gov.ru/news/page/transportnyj-kompleks-vhodit-v-paterku-krupnejsih-rossijskih-otraslej-28261 (дата обращения: 25.06.2026).

[2] Wang Zh., Li P., Li Q., Wang Zh., Li Zh. Motion planning method for car-like autonomous mobile robots in dynamic obstacle environments // IEEE Access. – 2023. – vol. 11. – pp. 137387–137400.

[3] Fraichard Th., Garnier Ph. Fuzzy control to drive car-like vehicles // Robotics and Autonomous Systems. – vol. 34. – iss. 1. – 2001. – pp. 1–22.

[4] Zhang R., Guo H., Sotelo M. A., Du H., Darius A., Li Z. New RRT-based method for vehicle path planning in curve scenarios considering path oscillations // IEEE Transactions on Vehicular Technology. – 2025. – vol. 74. – no. 12. – pp. 18445–18459.

[5] Farag W. Kalman-filter-based sensor fusion applied to road-objects detection and tracking for autonomous vehicles // Proc. of the Institution of Mechanical Engineers, Part I: Journal of Systems and Control Engineering. – 2020. – Vol. 235. – no. 3.– pp. 095965182097552.

[6] Chen W., Chi W., Ji S., Ye H., Liu J., Jia Y., Yu J., Cheng J. A survey of autonomous robots and multi-robot navigation: Perception, planning and collaboration //Biomimetic Intelligence and Robotics. – 2025. – vol. 5. – no. 2. – pp. 100203.

[7] Kebbati Y., Oufroukh N. A., Vigneron V., Ichalal D., Gruyer D. Optimized self-adaptive PID speed control for autonomous vehicles // 26th International Conference on Automation and Computing (ICAC 2021). – 2021. – pp. 1–6.

[8] Ofoegbu E. Artificial intelligence and machine learning in autonomous driving: current trends and future directions: a review // J. Eng. Appl. Sci. – 2026. – vol. 73. – no. 251.

[9] Архиерейский М.М., Диане С.А.К. Мониторинг и прогнозирование состояний электроприводов колесного робота // Системная инженерия и инфокоммуникации. – 2025. – № 3. – С. 4–8.

[10] Кутузов Д.В., Осовский А.В., Мальцева Н.С., Мартынов С.В. Свойства трафика видеоконференций 5G и его прогнозирование методами искусственного интеллекта // Вестник Астраханского государственного технического университета. Серия: Управление, вычислительная техника и информатика. – 2025. – no. 1. – pp. 103–116.

[11] Fox D., Burgard W., Thrun S. The dynamic window approach to collision avoidance // IEEE Robotics & Automation Magazine. – 1997. – V. 4 (1). – pp. 23–33.

[12] Zhang Y., Xiao Zh., Yuan X., Li S., Liang S. Obstacle avoidance of two-wheeled mobile robot based on DWA algorithm. – 2019. – pp. 5701–5706.

[13] Van Den Berg J., Lin M., Manocha D. Reciprocal n-body collision avoidance // Robotics Research: The 14th International Symposium ISRR. – Berlin, Heidelberg: Springer Berlin Heidelberg. – 2011. – pp. 3–19.

[14] Kien D., Quoc Phong P., Hai N., Thuong L., Hao H. Study on kinematic of two-wheeled and four-wheeled steering systems // Journal of science and development economics. – 2026. – no. 1. – pp. 79–91.

Информация об авторах

Диане Секу Абдель Кадер, канд. техн. наук, старший научный сотрудник ИПУ РАН, доцент кафедры проблем управления РТУ МИРЭА, Москва, Россия, e-mail: diane1990@yandex.ru, ORCID: 0000-0002-8690-6422

Назаркин Владимир Иванович, аспирант РТУ МИРЭА, Москва, Россия, nazarkin.v.i@edu.mirea.ru

Бессонова Александра Викторовна, аспирант РТУ МИРЭА, Москва, Россия, bessonova@mirea.ru

Суслова Анастасия Владиславовна, магистрант РТУ МИРЭА, suslova.a.v2@edu.mirea.ru

Predictive control algorithm for an autonomous vehicle in an environment with obstacles

Sekou A. K. Diane [1,2], Vladimir I. Nazarkin [2], Alexandra V. Bessonova [2], Anastasia V. Suslova [2]

[1]V.A. Trapeznikov Institute of Control Sciences, Moscow, Russia
[2]MIREA – Russian Technological University, Moscow, Russia

***Abstract* – The key tasks in the operation of an autonomous vehicle are known to be following the intended route and avoiding collisions with obstacles. In this paper, a control algorithm is proposed based on a predictive simulation model of an unmanned platform, which allows for the implementation of such predictive behavior. The methodological framework includes the use of the Euler method for numerically solving the vehicle's kinematics differential equations, the dynamic window method for evaluating and selecting possible vehicle trajectories under different control decisions, and the reduced fuzzy logic inference method for selecting route target points. Experiments conducted in a virtual environment confirm the effectiveness of the approach in test motion scenarios.**

***Keywords* – predictive control, obstacle avoidance, dynamic window method, autonomous transport, fuzzy logic.**

REFERENCES

[1] The transport complex is one of the five largest Russian industries // Analytical Center for the Government of the Russian Federation. URL: https://ac.gov.ru/news/page/ transportnyj-kompleks-vhodit-v-paterku-krupnejsih-rossijskih-otraslej-28261 (accessed on 25.06.2026) (In Russ.).

[2] Wang Zh., Li P., Li Q., Wang Zh., Li Zh. Motion planning method for car-like autonomous mobile robots in dynamic obstacle environments // IEEE Access. – 2023. – vol. 11. – pp. 137387–137400.

[3] Fraichard Th., Garnier Ph. Fuzzy control to drive car-like vehicles // Robotics and Autonomous Systems. – vol. 34. – iss. 1. – 2001. – pp. 1–22.

[4] Zhang R., Guo H., Sotelo M. A., Du H., Darius A., Li Z. New RRT-based method for vehicle path planning in curve scenarios considering path oscillations // IEEE Transactions on Vehicular Technology. – 2025. – vol. 74. – no. 12. – pp. 18445–18459.

[5] Farag W. Kalman-filter-based sensor fusion applied to road-objects detection and tracking for autonomous vehicles // Proc. of the Institution of Mechanical Engineers, Part I: Journal of Systems and Control Engineering. – 2020. – Vol. 235. – no. 3. – pp. 095965182097552.

[6] Chen W., Chi W., Ji S., Ye H., Liu J., Jia Y., Yu J., Cheng J. A survey of autonomous robots and multi-robot navigation: Perception, planning and collaboration //Biomimetic Intelligence and Robotics. – 2025. – vol. 5. – no. 2. – pp. 100203.

[7] Kebbati Y., Oufroukh N. A., Vigneron V., Ichalal D., Gruyer D. Optimized self-adaptive PID speed control for autonomous vehicles // 26th International Conference on Automation and Computing (ICAC 2021). – 2021. – pp. 1–6.

[8] Ofoegbu E. Artificial intelligence and machine learning in autonomous driving: current trends and future directions: a review // J. Eng. Appl. Sci. – 2026. – vol. 73. – no. 251.

[9] Archiereiskiy M.M., Diane S.A.K. Monitoring and predicting the states of electric drives of a wheeled robot // System Engineering and Infocommunications. – 2025. – No. 3. – pp. 4–8. (In Russ.)

[10] Kutuzov D. V., Osovsky A. V., Maltseva N. S., Martynov S. V. Properties of 5G video conferencing traffic and its prediction by artificial intelligence methods // Bulletin of Astrakhan State Technical University. Series: Management, Computer Science, and Informatics. – 2025. – No. 1. – pp. 103–116. (In Russ.)

[11] Fox D., Burgard W., Thrun S. The dynamic window approach to collision avoidance // IEEE Robotics & Automation Magazine. – 1997. – V. 4 (1). – pp. 23–33.

[12] Zhang Y., Xiao Zh., Yuan X., Li S., Liang S. Obstacle avoidance of two-wheeled mobile robot based on DWA algorithm. – 2019. – pp. 5701–5706.

[13] Van Den Berg J., Lin M., Manocha D. Reciprocal n-body collision avoidance // Robotics Research: The 14th International Symposium ISRR. – Berlin, Heidelberg: Springer Berlin Heidelberg. – 2011. – pp. 3–19.

[14] Kien D., Quoc Phong P., Hai N., Thuong L., Hao H. Study on kinematic of two-wheeled and four-wheeled steering systems // Journal of science and development economics. – 2026. – no. 1. – pp. 79–91.

Information about the authors

Sekou Abdel Kader Diane, Ph.D. (Eng.), Senior Researcher at the Institute of Control Sciences RAS, Associate Professor at the Department of Problems of Control at RTU MIREA, Moscow, Russia, e-mail: diane1990@yandex.ru, ORCID: 0000-0002-8690-6422

Vladimir Ivanovich Nazarkin, Postgraduate Student, RTU MIREA, Moscow, Russia, nazarkin.v.i@edu.mirea.ru

Alexandra Viktorovna Bessonova, Postgraduate Student, RTU MIREA, Moscow, Russia, bessonova@mirea.ru

Anastasia Vladislavovna Suslova, Master's Student, RTU MIREA, suslova.a.v2@edu.mirea.ru

Обнаружение зон аномальных изменений в многомерных рядах коммерческого учета тепловой энергии

О.В. Стукач [1,2], О.О. Еремина [3]

[1]*Новосибирский государственный технический университет, Новосибирск, Россия*
[2]*Национальный исследовательский университет «Высшая школа экономики», Москва, Россия*
[3]*Национальный исследовательский Томский государственный университет Томск, Россия*

Аннотация **– Обнаружение аномальных значений в таблицах данных, полученных с тепловычислителей, необходимо для повышения качества учёта потребления тепловой энергии. К зонам аномальных значений приводят поломки приборов учёта, ошибки передачи, человеческий фактор и множество других причин. И если обнаружение грубых ошибок измерений статистическими методами сейчас не вызывает особых сложностей, обнаружение точек аномальных изменений и резких изменений в рядах – это более сложная задача, которая привлекает внимание исследователей во многих предметных областях, где активно внедряется интеллектуальный анализ данных. В конструировании погодных регуляторов отопления многоэтажных жилых зданий одним из задающих воздействий является температура наружного воздуха с резкими перепадами, вызванными, главным образом, глобальным изменением климата. В статье предлагается использовать адаптивную состоятельную меру обнаружения аномальных изменений ряда, основанную на выборе функций, наименее уклоняющихся от многомерных рядов. Мера основана на различиях в распределении вероятностей до и после зон аномальных изменений. Предложенный подход обнаружения устойчив к погрешностям измерения и может быть использован для повышения точности оценки работы регуляторов тепловой энергии. На опубликованных датасетах, находящихся в открытом доступе, проведено тестирование разработанного алгоритма. Поскольку аномалии в рядах зачастую встречаются в данных автоматических систем контроля и передачи данных учёта энергии, для тестирования взяты распространенные датасеты. Разработанный алгоритм по времени анализа данных сопоставим с популярным алгоритмом SVM, а по F-статистике демонстрирует преимущество перед другими известными методами. Предложенный способ обнаружения зон изменения в рядах является многообещающим в задачах экономии энергии и продемонстрирован с помощью обширных вычислительных экспериментов на реальных наборах данных теплоучёта.**

Ключевые слова **– качество данных, оценка параметров, алгоритм обработки сигналов, идентификация системы управления, коммерческий учет потребления энергоресурсов.**

ВВЕДЕНИЕ

Задача анализа дискретных временных рядов с целью определения природы ряда и прогнозирования значений часто встречается на практике и достаточно хорошо исследована для рядов с разными особенностями поведения [1]. Разработаны многочисленные методы расчёта трендов, оценки периодичности, выделения компонент ряда и др [2]. Менее исследованы способы обнаружения точек и зон аномального поведения рядов, если ошибки или особые значения не превышают обычной дисперсии данных [3]. Особую актуальность приобретают методы анализа связанных рядов, или многомерных рядов [4, 5] и модели обнаружения связи между временными рядами [6]. Многие практические задачи в моделировании рядов приводят к необходимости не только прогнозирования, но и разного рода статистических оценок в пределах ряда [7]. Также важно учитывать ситуацию, когда источник данных нестационарный, поскольку характер источника данных может меняться с течением времени в реальных условиях. На практике достаточно сложно получить длинные ряды, а зашумленные зоны аномалий имеют большую продолжительность [8]. Известные алгоритмы поиска аномалий плохо справляются с их точным анализом при большой зашумленности и в условиях слабых статистических различий между измерениями, особенно когда их достаточно много.

Актуальной задачей анализа аномалий в рядах является коммерческий учёт тепловой энергии, где должно контролироваться качество данных, используемых в автоматизированных системах начислений платы за потреблённые энергоресурсы [9]. Зачастую некорректные показания приборов учёта не выявляются или выявляются не своевременно. Увеличению ошибок и снижению качества способствует интеграция данных из разных источников. Иногда данные просто не передаются на протяжении некоторого времени [10].

Аномалии скрыты шумом и обычными вариациями. Это характерно для многих данных, например учёта потреблённой электрической и тепловой энергии [11].

Для повышения качества данных в статье [12] предлагается двухэтапная очистка временных рядов показаний приборов учета: сначала применяется иерархическая кластеризация для выделения грубых ошибок, а затем в результате статистической обработки выявляются выбросы. Но окончательное решение об ошибочности наблюдений принимается экспертным путём. Поскольку полная автоматизация предполагает исключение экспертных оценок, необходимо разрабатывать новые способы выявления аномалий и последующей очистки данных.

I. ПОДХОДЫ К ОБНАРУЖЕНИЮ АНОМАЛИЙ

Известные подходы к обнаружению аномалий можно разделить на три класса: методики, основанные на подпространстве [13], на соседстве [14] и комбинированные [15]. Среди последних можно выделить способ оценки плотности аномалий – Local Outlier Factor (LOF), метод машинного обучения Support Vector Machine (SVM), Elliptic Envelope – это обнаружение аномалий по оценке многомерного нормального распределения, Subspace Outlier Detection (SOD) – метод кластеризации с выделением аномалий.

Многообещающим методом машинного обучения для обнаружения точек аномальных изменений для многомерных рядов является анализ стационарного подпространства (stationary subspace analysis, SSA) [16]. В этом методе многомерные ряды разлагаются на стационарные и нестационарные компоненты, а затем в нестационарном подпространстве могут быть обнаружены точки изменения. Поскольку метод SSA позволяет уменьшить размерность данных без потери зон аномальных изменений, он может значительно повысить скорость обнаружения. Недостатком метода SSA является то, что при малой размерности пространства значений рядов необходимо вычислять логарифм сингулярной ковариационной матрицы. Для точной факторизации рядов на компоненты требуется относительно большое число обучающих выборок. Можно с уверенностью считать, что метод SSA непригоден для анализа ограниченного объёма данных учёта тепловой энергии для выбранного теплосчётчика, так как число значений рядов равно числу дней отопительного сезона.

Идеология применения методов машинного обучения с учителем развита в работе [17], где ставится задача построения классификатора, дискриминирующая ряды, если они произошли от разных источников. Предложенный авторами алгоритм работает на синтетических данных с биномиальным распределением, но для реальных данных потребления тепловой энергии классификатор пропускает больше половины случаев аномалий.

Если предположить, что существует точка аномалии $q(t)$ в момент t, распределение вероятностей данных $q(t\text{-}1)$, $q(t\text{-}2)$,... в прошлом t-1, t-2... должно отличаться от распределения $q(t+1)$, $q(t+2)$,... в будущем t+1, t+2. Если это различие значимо, можно считать, что зона аномалии обнаружена. На основе этого предположения были предложены различные подходы к обнаружению точки изменения с кумулятивной суммой (likelihood ratio test) и на основе общеизвестного отношения правдоподобия [18]. Здесь логарифм отношения правдоподобия между двумя распределениями вероятностей используется в качестве меры обнаружения изменения, где каждая плотность вероятности оценивается независимо. Но поскольку сама по себе оценка плотности вероятности в условиях погрешностей измерения является непростой задачей, известные подходы работают плохо, особенно на реальных данных коммерческого учёта тепловой энергии [19]. Кроме того, методы оценки плотности вероятности в зависимых рядах довольно сложны и непрактичны для применения в регуляторах, так как требуют сложных компьютерных вычислений, а результат получается весьма приближённым [20]. К тому же регулятор может отработать такую ошибку без потери устойчивости.

Чтобы избежать использования оценки плотности вероятности, были предложены прямые подходы к определению зоны изменения, основанные на состоятельных мерах зависимости [21, 22], а также оценке соотношения вероятностных распределений непосредственно без использования оценки плотности [23, 24] или методами машинного обучения [25]. Они дают оптимальную скорость работы алгоритмов, но относительно низкую точность из-за влияния погрешности измерений.

II. ПОСТАНОВКА ЗАДАЧИ

Перспективным представляется подход, использующий критерий независимости Гильберта-Шмидта (КНГШ) на основе ядра [26]. КНГШ работает на основе выборок объекта и соответствующих ему бинарных переменных +1 (будущие данные) и –1 (прошлые данные) с разными распределениями вероятностей. Преимущество оценки КНГШ по сравнению с рассмотренными ранее мерами обнаружения заключается в том, что он может включать выбор признака во время оценки меры обнаружения. Сначала выбираются значения, которые отвечают за резкое изменение ряда, а затем вычисляется оценка КНГШ, используя эти выбранные значения. Зашумленность данных оказывает меньшее влияние на результат, чем рассмотренные методы, и предлагаемый метод перспективен в рассматриваемой предметной области.

Пусть $q(i)$ – m-размерная выборка ряда в момент i. Если выборка сделана из одного и того же

распределения, – в ряде аномального изменения нет, в противном случае есть. Методом скользящего окна в момент i из $q(i)$ извлекаются n-мерные выборки $q(i-j+1)\big|_{j=1}^{n}$. Две непересекающиеся последовательности $q(t)$ и $q(t+n)$ отметим как $r(i)=1$ и $r(i+n)=-1$ соответственно. Рассмотрим объединённый ряд $Q(i):\{q(i+n-j+1)\}\big|_{j=1}^{2n}$, для которого $r(i)=\pm 1$. Обнаружение зон аномального изменения ряда заключается в вычислении меры M зависимости между q и r из $Q(i)$, такой, что

$$\begin{cases} Q(i) < s \Rightarrow \text{изменений нет;} \\ Q(i) \geq s \Rightarrow \text{аномалия обнаружена,} \end{cases}$$

где s – порог обнаружения зоны аномалии. Поскольку данные, как правило, измеряются с большими погрешностями и зашумлены, факт обнаружения сильно зависит от s.

Представим КНГШ $H[.]$ в качестве меры зависимости $Q(i)$ в виде [23]:

$$Q(i)=\sum_{k=1}^{n}\xi_k H[B_k(i),C],$$

где $\xi_k>0$, $B(i)=\begin{bmatrix} q(i-n+1) \\ q(i-n+2) \\ \dots \\ q((i+n-1)) \end{bmatrix}$, $C=\begin{bmatrix} 1 \\ 1 \\ \dots \\ -1 \end{bmatrix}$,

$H[B_k(i),C]=tr(D_k^{\text{н}}E^{\text{н}})$, $D^{\text{н}}=\Gamma D\Gamma$, $E^{\text{н}}=\Gamma E\Gamma$, $D^{\text{н}}$, $E^{\text{н}}$ – нормализованные центрированные матрицы ядра, Г – матрица Грама.

КНГШ является состоятельной мерой зависимости случайных величин x и y, удовлетворяющей условию $0\leq H[x,y]\leq 1$ и равен нулю тогда и только тогда, когда x и y статистически независимы. Если используется гауссово ядро, $H[x,y]\to 0$ [27].

При разработке эвристических алгоритмов поиска аномалий можно принять $\xi_k\equiv 1$. Но для автоматического поиска аномалий целесообразно использовать какой-либо регуляризующий метод для поиска ξ_k, поскольку равные весовые коэффициенты оценок КНГШ не являются хорошим выбором для многомерных рядов, особенно при зашумленных данных, например

$$\min_{\xi}\left\|E^{\text{Н}}-\sum_{\text{k}}^{\text{n}}\xi_k D_k^{\text{Н}}\right\|^2+\beta\|\xi\|,\ \xi\big|_{k=1}^{n}\geq 0,$$

где β – параметр регуляризации.

Здесь матрица ядра $E^{\text{н}}$ аппроксимируется линейной функцией элементов матрицы $D_k^{\text{н}}$. После нахождения ξ_k эти элементы нормализуются. Поскольку

$$\min_{\xi}\left\|E^{\text{Н}}-\sum_{\text{k}}^{\text{n}}\xi_k D_k^{\text{Н}}\right\|^2=H[C,C]- \\ -2\sum_{k=1}^{n}\xi_k H[B_k,C]+\sum_{i,j}^{n}\xi_i\xi_j H[B_iB_j],$$

а первое и третье слагаемые равны нулю ввиду взаимной независимости, для минимизации целевой функции необходимо найти максимум $Q(i)=\sum_{k=1}^{n}\xi_k H[B_k(i),C]$. В некоторой степени этот вывод очевиден, ведь если k-е наблюдение $B_k(i)$ имеет сильную зависимость от C, КНГШ $H[B_k(i),C]$ принимает большое значение и, следовательно, это наблюдение сохраняется в процессе регуляризации. Если же $B_k(i)$ и C взаимно независимы, КНГШ стремится к нулю, и такое наблюдение вскоре удаляется. В результате остаются только те наблюдения, которые приводят к резкому изменению результирующей переменной. Избыточные значения, соответствующие слагаемым $\xi_i\xi_j H[B_iB_j]$, также удаляются в процессе регуляризации.

III. АПРИОРНАЯ ИНФОРМАЦИЯ О ФУНКЦИЯХ

Если имеется предварительная информация о свойствах рядов, её можно использовать для выбора функций в КНГШ. Например, при наличии в базе данных нескольких энергоэффективных домов резкое изменение теплопотребления происходит в каждом из них в близкий момент времени [28]. Поэтому для повышения эффективности регуляризации вполне разумно выбрать ряды, соответствующие кластеру энергоэффективных домов, исходя из каких-то предварительных предположений. Этому может помочь знание города [29], и с этой целью можно использовать регуляризатор группового типа эластичной сети:

$$\min_{\xi}\frac{1}{2}\left\|E^{\text{Н}}-\sum_{\text{k}}^{\text{n}}\xi_k D_k^{\text{Н}}\right\|^2+\mu\sum_{l=1}^{L}\|\gamma_l\| \qquad (1)$$

где $\gamma=[\gamma_1,\gamma_2,\dots\gamma_L]^{\text{T}}$ – l-я группа переменных, L – количество групп, μ – параметр регуляризации. Для

минимизации (1) предлагается следующий пошаговый алгоритм (Рис. 1).

1 Пусть $D^{\text{н}} \in R$ – вектор начальных приближений

2 $D_0^{\text{н}} \leftarrow \min \sum_{l=1}^{L} \|\gamma_l\|$

3 Цикл l=1, L

Если $\left\| E^{\text{н}} - \sum_{k}^{n} \xi_k D_k^{\text{н}} \right\|^2 \le \mu \sum_{l=1}^{L} \|\gamma_l\|$

$\zeta_l \leftarrow 0$

иначе $\zeta \leftarrow \min \sum_{l=1}^{L} \|\gamma_l\|$

конец цикла

4 Повторить шаги 2, 3 до достижения критерия

Рис. 1. Алгоритм регуляризации

IV. ВЫБОР ЯДРА

КНГШ изменяется в зависимости от параметров и вида ядра. Поэтому входные параметры ядра для всех функций должны быть фиксированы. Сначала проводится нормализация характеристик q к распределению с нулевым средним q_0 и стандартным отклонением σ, равным единице. Затем для всех ядер используются одни и те же параметры. Для входных переменных q используется гауссово ядро: $K(q,q_0) = \exp[\frac{(q-q_0)^2}{2\sigma^2}]$, где $\sigma = 1$. Для переменной r используется δ-ядро, $K(r,r_0) = \begin{cases} 1, r = r_0 \\ 0, r \neq r_0 \end{cases}$.

V. ЧИСЛЕННЫЙ ЭКСПЕРИМЕНТ С НАБОРОМ СИНТЕТИЧЕСКИХ ДАННЫХ

Для проверки работоспособности алгоритма были выбраны датасеты из открытых источников [28, 30–35]. Для КНГШ использовалось гауссово ядро, параметр регуляризации $\mu = 0{,}1$.

Для оценки точности C и полноты R предсказания аномалий используется F-метрика, известная из задач классификации:

$$F = 2CR/(C+R), F \in [0,1],$$

где $C = TP/(TP+FP)$, $R = FP/(TN+FN)$, а значения параметров TP (True Positive), FP (False Positive), TN (True Negative), FN (False Negative) соответствуют Табл. I.

ТАБЛИЦА I
F-МЕТРИКИ

Параметры	Действительное наличие аномалии	Предсказание аномалии
TP – правильно определённые аномалии	1	1
TN – правильно определённые нормальные данные	0	0
FP – правильные значения, которые неправильно определены как аномальные	0	1
FN – аномалии, которые определены как правильные значения	1	0

Известно, что чем ближе значение F к единице, тем больше вероятность того, что будут корректно обнаружены аномалии в данных. Результаты тестирования представлены на Рис. 2.

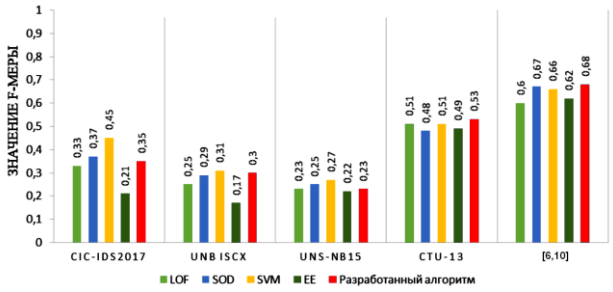


Рис. 2. Значения F-мер

Из Рис. 2 видно, что по F-оценке разработанный алгоритм в среднем обнаруживает аномалии чуть менее точно, чем алгоритм SVM, но по крайней мере в 1,1 раз лучше, чем LOF.

Кроме точности обнаружения аномалий в режиме реального времени важна оценка скорости работы алгоритмов. Производительность оценивалась по времени выполнения программы до полного обнаружения аномалий (Рис. 3).

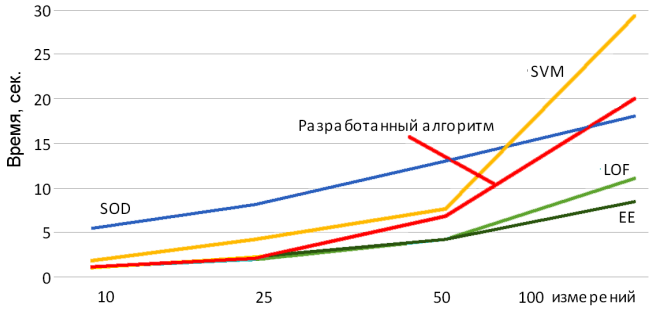


Рис. 3. Зависимости времени анализа данных от количества измерений

Из графиков на Рис. 3 видно, что разработанный алгоритм в целом сопоставим по производительности с другими алгоритмами поиска аномалий, и наряду с

ними может использоваться при исследовании многомерных рядов данных.

VI. ОБСУЖДЕНИЕ

Из проведённого тестирования на опубликованных размеченных данных из открытых источников с указанием мест аномалий [30–35] можно сделать следующие выводы.

1) Алгоритм имеет преимущество в производительности по сравнению с алгоритмом SVM, который по рассчитанной *F*-метрике сравним с разработанным.

2) Крайне неэффективным показал себя алгоритм Elliptic Envelope. По-видимому, он совсем непригоден для выявления аномалий. Его точность почти вполовину ниже, чем у SVM и предлагаемого алгоритма.

3) Предложенный алгоритм вполне может конкурировать с другими популярными подходами к обнаружению аномалий, как по точности обнаружения, так и по времени анализа данных.

ЗАКЛЮЧЕНИЕ

Обнаружение зон аномальных изменений в многомерных рядах привлекает значительное внимание в области интеллектуального анализа данных. Эта задача имеет большое количество практических приложений, в частности, повышение качества учёта потребления тепловой энергии.

В статье исследован подход к обнаружению аномальных изменений в рядах на основе [17], использующий критерий независимости Гильберта-Шмидта (КНГШ) на основе ядра [26], и который работает на основе выборок и соответствующих ему бинарных данных с разными распределениями вероятностей. Преимущество предлагаемого подхода перед существующими заключается в том, что оценка наличия аномалии проводится с использованием признаков, которые важны для резких изменений. Предлагаемый подход более устойчив к зашумлению данных, чем существующие методы. Проведением численных экспериментов с синтетическими и реальными наборами данных показана перспективность предлагаемого подхода к обнаружению зон аномальных изменений.

ЛИТЕРАТУРА

[1] Крюков Ю.А., Чернягин Д.В. ARIMA – модель прогнозирования значений трафика // Информационные технологии и вычислительные системы. – 2011. – № 2. – С. 41-49. – eLIBRARY ID: 17805262. – EDN: OZQLLR.

[2] Газиров Д. И. Обзор методов статистического анализа временных рядов и проблемы, возникающие при анализе нестационарных временных рядов // Научный журнал: Олимп (Иваново). – 2016. – № 3(4). –С. 9-14. – eLIBRARY ID: 25631242.

[3] Zhang W., Li T., He P., Yang Y., Wang S. An Outlier Suppression and Adversarial Learning Model for Anomaly Detection in Multivariate Time Series // Entropy. – 2025. – Vol. 27. – № 11. – P. 1151. – Doi: 10.3390/e27111151.

[4] Swami A., Giannakis G., Shamsunder S. Multichannel ARMA processes // IEEE Transactions on Signal Processing. – 1994. – Vol. 42. – № 4. – Pp. 898-913. – Doi: 10.1109/78.285653.

[5] Ivanovic M., Kurbalija V. Time series analysis and possible applications / 2016 39th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO). – Opatija, Croatia, May 30–June 3, 2016. – Pp. 473-479. – Doi: 10.1109/MIPRO.2016.7522190.

[6] Усманова К.Р., Стрижов В.В. Модели обнаружения зависимостей во временных рядах в задачах построения прогностических моделей // Системы и средства информатики. – 2019. – Т. 29. – № 2. – С. 12-30. – eLIBRARY ID: 41273551.

[7] Осовский А.В., Кутузов Д.В., Стукач О.В. Анализ моделей трафика, создаваемого устройствами интернета вещей / Динамика систем, механизмов и машин. – 2019. – Т. 7. – № 4. – С. 220-226. – ISSN: 2310-9793. – eLIBRARY ID: 41432222.

[8] Кузнецов М.П., Ивкин Н.П. Алгоритм классификации временных рядов акселерометра по комбинированному признаковому описанию // Машинное обучение и анализ данных. – 2015. – Т. 1. – № 11. – С. 1471-1483. – eLIBRARY ID: 24931576.

[9] Савкина А.В., Федосин А.С. Проблемы качества данных в автоматизированных системах коммерческого учета потребления энергоресурсов // Прикаспийский журнал: управление и высокие технологии. – 2014. – № 2. – С. 158-164. – eLIBRARY ID: 21716788. – EDN: SHFYGT.

[10] Стукач О.В., Габитова О.О. Демонстрация результатов энергосберегающего регулирования тепловой энергии на примере одного здания // Автоматика и программная инженерия. – 2023. – № 2(44). – Pp. 158-164. – eLIBRARY ID: 54207953.

[11] Попов И.Ю., Стукач О.В., Зорин П.А. Оценка качества данных коммерческого учёта тепловой энергии методом полной вариации распознавания ошибок // Динамика систем, механизмов и машин. – 2021. – Т. 9. – № 3. – С. 117-121. – DOI: 10.25206/2310-9793-9-3-117-121.

[12] Федосин А.С., Федосин С.А. Очистка входных данных в автоматизированных системах контроля и учета энергоресурсов // Инфокоммуникационные технологии. – 2016. – Т. 14. – № 2. – С. 162-168.

[13] Bacher M., Ben-Gal I., Shmueli E. Subspace selection for anomaly detection: An information theory approach. / IEEE International Conference on the Science of Electrical Engineering (ICSEE). – 16–18 Nov. 2016. – Eilat. – Israel, id.50. – P. 1-5. – Doi: 10.1109/icsee.2016.7806077.

[14] Dutta J.K., Banerjee B., Reddy C.K. RODS: Rarity based Outlier Detection in a Sparse Coding Framework // IEEE Transactions on Knowledge and Data Engineering, – 2016. – Vol. 28. – № 2. – Pp. 483-495. – Doi: 10.1109/TKDE.2015.2475748.

[15] Shebuti R. Sequential Ensemble Learning for Outlier Detection: A Bias-Variance Perspective // 2016 IEEE 16th International Conference on Data Mining (ICDM). – 2016. – Pp. 1167–1172. – arXiv:1609.05528. – doi: 10.48550/arXiv.1609.05528.

[16] Blythe D., Bunau P., Meinecke F.C., Muller K.R. Feature Extraction for Change-Point Detection using Stationary Subspace Analysis // IEEE Transactions on Neural Networks and Learning Systems. – 2012. – Vol. 23. – № 4. – Pp. 631–643. – arXiv:1108.2486v1. – Doi: 10.1109/TNNLS.2012.2185811.

[17] Hido S., Ide T., Kashima H., Kubo H., Matsuzawa H. Unsupervised change analysis using supervised learning / Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD). – 2008. – Part of the book series: Lecture Notes in Computer Science (LNAI, vol. 5012). – Pp. 148–159.

[18] Gustafsson F. The marginalized likelihood ratio test for detecting abrupt changes // IEEE Transactions on Automatic Control. – 1996. Vol. 41. – № 1. – Pp. 66–78. – Doi: 10.1109/9.481608.

[19] Попов И.Ю., Стукач О.В., Зорин П.А. Применение метода полной вариации распознавания ошибок к исследованию

данных коммерческого учета тепловой энергии // Автоматика и программная инженерия. – 2021. – № 4(38). – С. 55-61.
[20] Лагутин М.В. Наглядная математическая статистика. – М.: ВИНОМ. Лаборатория знаний, 2009. – 472 с.
[21] Ershov I.A., Stukach O.V. Aimagambetova R.Z. The Peculiar Measure Identifying of the Temperature Leap in the Distributed Raman Sensors // Dynamics of Systems, Mechanisms and Machines (Dynamics). – 2021, 9-11 Nov., Omsk, Russian Federation. – Pp. 1-4. – Doi: 10.1109/Dynamics52735.2021.9653721.
[22] Ершов И.А., Стукач О.В., Цыденжапов И.Б., Сычев И.В. Идентификация слабого температурного перехода в оптоволокне при помощи дивергенции Цаллиса / Электронные средства и системы управления: материалы докладов XVI Международной научно-практической конференции (18-20 ноября 2020 г.): в 2 ч. – Ч. 1. – Томск: В-Спектр. – 2020. – С. 146-148.
[23] Nguyen X.L., Wainwright M.J., Jordan M.I. Estimating divergence functionals and the likelihood ratio by convex risk minimization // IEEE Transactions on Information Theory. – 2010. – Vol. 56. – № 11. – Pp: 5847–5861. – arXiv:0809.0853v2. – Doi: 10.1109/TIT.2010.2068870.
[24] Kawahara Y., Sugiyama M. Change-point detection in time-series data by direct density-ratio estimation / SIAM International Conference on Data Mining (SDM 2009). – April 30 - May 2. – 2009. – Sparks, Nevada, USA. – Pp. 389-400. – Doi:10.1137/1.9781611972795.34.
[25] Liu S., Yamada M., Collier N., Sugiyama M. Change-point detection in time-series data by relative density ratio estimation // Neural Networks. – 2013. – 16 Jan. – arXiv:1203.0453v2. – Doi: 10.48550/arXiv.1203.0453.
[26] Gretton A., Bousquet O., Smola A., Scholkopf B. Measuring statistical dependence with Hilbert-Schmidt norms // ALT papes. – 2005. – Pp. 63-77.
[27] Steinwart I. On the influence of the kernel on the consistency of support vector machines // Journal of Machine Learning Research. – 2001. – № 2. – Pp. 67-93.
[28] Zorin P., Stukach O. Data of heating meters from residential buildings in Tomsk (Russia) for statistical modeling of the thermal characteristics of buildings / IEEE Dataport. – 2020. – [Online]. Available: http://dx.doi.org/10.21227/3r4e-ch18.
[29] Зорин П.А., Стукач О.В. Анализ влияния погодных условий на динамику тепловой энергии в жилом фонде города Томск / Инновационные, информационные и коммуникационные технологии: сборник трудов XVII Международной научно-практической конференции. – М.: Ассоциация выпускников и сотрудников ВВИА им. проф. Жуковского. – 2020. – С. 365-368. – eLIBRARY ID: 45557868.
[30] Набор данных CIC-IDS2017. Режим доступа: https://www.unb.ca/cic/datasets/ids-2017.html.
[31] Stukach O., Zorin P. Long-Term Data from the Heat Meters in Residential Buildings Depending on the Outside Temperature and Characteristics of Buildings / IEEE Dataport. – April 13, 2021. – Doi: 10.21227/cw53-rr81.
[32] Набор данных UNB ISCX. Режим доступа: https://www.unb.ca/cic/datasets/ids.html.
[33] Набор данных UNS-NB15. Режим доступа: https://research.unsw.edu.au/projects/unsw-nb15-dataset.
[34] Набор данных CTU-13. Режим доступа: https://www.kaggle.com/datasets/doogla/ctu13.
[35] Зорин П.А., Стукач О.В. База данных потребления тепловой энергии многоэтажными жилыми зданиями в зависимости от метеорологических факторов и характеристик зданий. – Рег. № 25452. – Doi: 10.12731/ofernio.2025.25452.

Информация об авторах

Стукач Олег Владимирович, д.т.н., профессор кафедры Защиты информации Новосибирского государственного технического университета, г. Новосибирск, Россия, профессор департамента Электронной инженерии Московского института электроники и математики им. А.Н. Тихонова Национального исследовательского университета «Высшая школа экономики», г. Москва, Россия, e-mail: tomsk@ieee.org, ORCID: 0000-0001-6845-4285.

Еремина Олеся Олеговна, студент Высшей IT-школы Национального исследовательского Томского государственного университета.

Detection of zones of abnormal changes in multidimensional series of commercial thermal energy metering

Oleg Stukach [1,2], Olesya Eremina[3]

[1]National Research University Higher School of Economics, Moscow, Russia
[2]Novosibirsk State Technical University, Novosibirsk, Russia
[3]National Research Tomsk State University, Tomsk, Russia

***Abstract* – The detection of abnormal values in the data tables from heat meters is necessary to improve the quality of accounting for heat energy consumption. Meter failures, transmission errors, human factors, etc. lead to abnormal value zones. Unless the detection of rough measurement errors by statistical methods is not particularly difficult now, the detection of points of abnormal changes and abrupt changes in the series is a more difficult problem. It attracts the attention of researchers in many fields where data analysis is actively being implemented. In the design of weather regulators for heating multi-storey residential buildings, one of the input is the temperature of the outside air with sudden fluctuations mainly follows from global climate change. We suggest the use an adaptive consistent measure for detecting abnormal series changes based on the selection of functions that least deviate from multidimensional series. The measure is based on differences in the probability distribution before and after the zones of abnormal changes. The proposed detection is resistant to measurement errors and can be used to improve the accuracy of thermal energy regulators. The designed algorithm has been tested on open datasets. One of the most common sets was used for testing because anomalies in the series are often found in the data of automatic systems for monitoring and transmitting energy metering data. The designed algorithm was comparable to the popular SVM algorithm in terms of data analysis time. It demonstrates an advantage over other well-known methods in terms of F-statistics. The proposed method for detecting zones of variation in series is promising in energy-saving tasks and has been demonstrated using extensive computational experiments on real heat accounting data sets.**

***Keywords* – data quality, parameter estimation, algorithm of signal processing, system identification, energy resources accounting.**

REFERENCES

[1] Kryukov Yu.A., Chernyagin D.V. An ARIMA model for forecasting a values of network traffic // Informatsionnie tekhnologii i vichislitelnie sistemi. – 2011. – № 2. – Pp. 41-49. – eLIBRARY ID: 17805262. – EDN: OZQLLR (in Russian).

[2] Gazirov D. I. Review of methods of statistical analysis of time series and problems arising in the analysis of non-stationary time series // Scientific journal: Olympus (Ivanovo). – 2016. – № 3(4). – Pp. 9-14. – eLibrary ID: 25631242 (in Russian).

[3] Zhang W., Li T., He P., Yang Y., Wang S. An Outlier Suppression and Adversarial Learning Model for Anomaly Detection in Multivariate Time Series // Entropy. – 2025. – Vol. 27. – № 11. – P. 1151. – Doi: 10.3390/e27111151.

[4] Swami A., Giannakis G., Shamsunder S. Multichannel ARMA processes // IEEE Transactions on Signal Processing. – 1994. – Vol. 42. – № 4. – Pp. 898-913. – Doi: 10.1109/78.285653.

[5] Ivanovic M., Kurbalija V. Time series analysis and possible applications / 2016 39th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO). – Opatija, Croatia, May 30–June 3, 2016. – Pp. 473-479. – Doi: 10.1109/MIPRO.2016.7522190.

[6] Usmanova K.R., Strizhov V.V. Models for detecting dependencies in time series in the tasks of constructing predictive models // Computer science systems and tools. – 2019. – Vol. 29. – № 2. – Pp. 12-30. – eLIBRARY ID: 41273551 (in Russian).

[7] Osovskiy A.V., Kutuzov D.V., Stukach O.V. Analyze traffic patterns generated by IoT devices // Dynamics of Systems, Mechanisms and Machines. – 2019. – Vol. 7. – № 4. – Pp. 220-226. – ISSN: 2310-9793. – eLIBRARY ID: 41432222 (in Russian).

[8] Kuznetsov M.P., Ivkin N.P. Time series classification algorithm using combined feature description // Mashinnoe obuchenie i analiz dannykh. – 2015. – Vol. 1. – № 11. – Pp. 1471-1483. – eLIBRARY ID: 24931576 (in Russian).

[9] Savkina A.V., Fedosin A.S. Data quality issues for energy management systems // Prikaspiisky zhurnal: upravlenie i visokie tekhnologii. – 2014. – № 2. – Pp. 158-164. – eLIBRARY ID: 21716788. – EDN: SHFYGT (in Russian).

[10] Stukach O.V., Gabitova O.O. Outcomes of Energy-Saving Regulation of Thermal Energy on the Example of One Building // Automatics & Software Enginery. – 2023. – № 2(44). – Pp. 158-164. – eLIBRARY ID: 54207953 (in Russian).

[11] Popov I.Yu., Stukach O.V., Zorin P.A. Evaluation of the Quality of Commercial Thermal Energy Accounting Data by Total Variation Outlier Recognizer // Dynamics of Systems, Mechanisms and Machines. – 2021. – Vol. 9. – № 3. – Pp. 117-121. – Doi: 10.25206/2310-9793-9-3-117-121 (in Russian).

[12] Fedosin A.S., Fedosin S.A. Input data cleaning in automatic systems for commercial measurement of power consumption // Infokommunikacionnye tekhnologii. – 2016. – Vol. 14. – № 2. – Pp. 162-168 (in Russian).

[13] Bacher M., Ben-Gal I., Shmueli E. Subspace selection for anomaly detection: An information theory approach. / IEEE International Conference on the Science of Electrical Engineering (ICSEE). – 16–18 Nov. 2016. – Eilat. – Israel, id.50. – P. 1-5. – Doi: 10.1109/icsee.2016.7806077.

[14] Dutta J.K., Banerjee B., Reddy C.K. RODS: Rarity based Outlier Detection in a Sparse Coding Framework // IEEE Transactions on Knowledge and Data Engineering, – 2016. – Vol. 28. – № 2. – Pp. 483-495. – Doi: 10.1109/TKDE.2015.2475748.

[15] Shebuti R. Sequential Ensemble Learning for Outlier Detection: A Bias-Variance Perspective // 2016 IEEE 16th International Conference on Data Mining (ICDM). – 2016. – Pp. 1167–1172. – arXiv:1609.05528. – doi: 10.48550/arXiv.1609.05528.

[16] Blythe D., Bunau P., Meinecke F.C., Muller K.R. Feature Extraction for Change-Point Detection using Stationary Subspace Analysis // IEEE Transactions on Neural Networks and Learning Systems. – 2012. – Vol. 23. – № 4. – Pp. 631–643. – arXiv:1108.2486v1. – Doi: 10.1109/TNNLS.2012.2185811.

[17] Hido S., Ide T., Kashima H., Kubo H., Matsuzawa H. Unsupervised change analysis using supervised learning / Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD). – 2008. – Part of the book series: Lecture Notes in Computer Science (LNAI, vol. 5012). – Pp. 148–159.

[18] Gustafsson F. The marginalized likelihood ratio test for detecting abrupt changes // IEEE Transactions on Automatic Control. – 1996. Vol. 41. – № 1. – Pp. 66–78. – Doi: 10.1109/9.481608.

[19] Popov I.Yu., Stukach O.V., Zorin P.A. The Use of Total Variation Outlier Recognizer Method to the Study of Commercial Thermal Energy Accounting Dataset // Automatics & Software Enginery. – 2021. – № 4(38). – Pp. 55-61 (in Russian).

[20] Lagutin M.V. Visual mathematical statistics. – Moscow: VINOM. Laboratory of Knowledge. – 2009. – 472 p. (in Russian).

[21] Ershov I.A., Stukach O.V. Aimagambetova R.Z. The Peculiar Measure Identifying of the Temperature Leap in the Distributed Raman Sensors // Dynamics of Systems, Mechanisms and Machines (Dynamics). – 2021, 9-11 Nov., Omsk, Russian Federation. – Pp. 1-4. – Doi: 10.1109/Dynamics52735.2021.9653721.

[22] Ershov I.A., Stukach O.V., Tsydenzhapov I.B., Sychev I.V. Identification of a weak temperature transition in optical fiber using the Tsallis divergence / Electronic tools and control systems: Proceedings of the XVI International Scientific and Practical Conference (November 18-20, 2020). – Part 1. – Tomsk. – V-Spektr, – 2020. – Pp. 146-148 (in Russian).

[23] Nguyen X.L., Wainwright M.J., Jordan M.I. Estimating divergence functionals and the likelihood ratio by convex risk minimization // IEEE Transactions on Information Theory. – 2010. – Vol. 56. – № 11. – Pp: 5847–5861. – arXiv:0809.0853v2. – Doi: 10.1109/TIT.2010.2068870.

[24] Kawahara Y., Sugiyama M. Change-point detection in time-series data by direct density-ratio estimation / SIAM International Conference on Data Mining (SDM 2009). – April 30 - May 2. – 2009. – Sparks, Nevada, USA. – Pp. 389-400. – Doi:10.1137/1.9781611972795.34.

[25] Liu S., Yamada M., Collier N., Sugiyama M. Change-point detection in time-series data by relative density ratio estimation // Neural Networks. – 2013. – 16 Jan. – arXiv:1203.0453v2. – Doi: 10.48550/arXiv.1203.0453.

[26] Gretton A., Bousquet O., Smola A., Scholkopf B. Measuring statistical dependence with Hilbert-Schmidt norms // ALT papes. – 2005. – Pp. 63-77.

[27] Steinwart I. On the influence of the kernel on the consistency of support vector machines // Journal of Machine Learning Research. – 2001. – № 2. – Pp. 67-93.

[28] Zorin P., Stukach O. Data of heating meters from residential buildings in Tomsk (Russia) for statistical modeling of the thermal characteristics of buildings / IEEE Dataport. – 2020. – [Online]. Available: http://dx.doi.org/10.21227/3r4e-ch18.

[29] Zorin P.A., Stukach O.V. Analiz vliyaniya pogodnykh uslovii na dinamiku teplovoi energii v zhilom fonde goroda Tomsk / Innovative, information and comms tech: Proceedings of the XVII Int. sci.-pract. conf. – Moscow. – 2020. – C. 365-368. – eLIBRARY ID: 45557868 (in Russian).

[30] The CIC-IDS2017 dataset. Access mode: https://www.unb.ca/cic/datasets/ids-2017.html.

[31] Stukach O., Zorin P. Long-Term Data from the Heat Meters in Residential Buildings Depending on the Outside Temperature and Characteristics of Buildings / IEEE Dataport. – April 13, 2021. – Doi: 10.21227/cw53-rr81.

[32] The UNB ISCX dataset. Access mode: https://www.unb.ca/cic/datasets/ids.html.

[33] The UNS-NB15 dataset. Access mode: https://research.unsw.edu.au/projects/unsw-nb15-dataset.

[34] The CTU-13 dataset. Access mode: https://www.kaggle.com/datasets/doogla/ctu13.

[35] Zorin P.A., Stukach O.V. Database of heat energy consumption by multi-storey residential buildings depending on meteorological factors and building characteristics. – Reg. № 25452. – Doi: 10.12731/ofernio.2025.25452 (in Russian).

Information about the authors

Oleg V. Stukach is with National Research University Higher School of Economics, Moscow, Russia; Novosibirsk State Technical University, Novosibirsk, Russia, e-mail: tomsk@ieee.org, ORCID: 0000-0001-6845-4285.

Olesya O. Eremina is with National Research Tomsk State University, Tomsk, Russia.

Применение общедоступных больших языковых моделей для проведения автоматизированного тестирования на проникновение

Д.А. Азява, Н.В. Иллак, А.А. Киселев

Новосибирский государственный технический университет, Новосибирск, Россия

Аннотация **– В работе исследуется возможность применения общедоступных больших языковых моделей для проведения кибератаки путем автоматизированного тестирования на проникновение в систему. Тестирование проводилось с использованием большой языковой модели DeepSeek-R1. Эксперимент был проведен на лабораторном стенде, состоящим из двух виртуальных машин с операционными системами: Kali Linux и Metasploitable2. Полученные результаты показали возможность применения большой языковой модели для проведения пентеста, но с некоторыми ограничениями.**

Ключевые слова **– БЯМ, ИИ, наступательный ИИ.**

ВВЕДЕНИЕ

Современный ландшафт киберугроз в начале 2026 года можно описать как динамический и сложный. Это обусловлено ростом количества обнаруживаемых уязвимостей (медианный прогноз некоммерческой организации по кибербезопасности FIRST составляет около 59 000 новых CVE [1]), а также качественной трансформацией инструментов атакующих. Ключевым драйвером этой трансформации становится тройственная природа технологий искусственного интеллекта (ИИ), где большие языковые модели (Large Language Model, LLM) выступают не только как механизм защиты или объект атак, но и как мощное средство для их проведения.

Несмотря на то, что уже существует множество исследований, направленных на изучение возможностей наступательного LLM [2-6], многие из этих исследований работают со специализированными LLM, написанными и обученными для обнаружения и эксплуатации уязвимостей. Однако, как показывает практика, серьёзную опасность несут и LLM общего назначения, находящиеся в общем доступе [7].

Наступательный потенциал LLM уже сегодня реализуется по нескольким ключевым направлениям. Уже сейчас LLM используется в сборе и обработке информации из открытых источников [8]. Также исследования показывают широкое применение LLM в методах социальной инженерии, это дает возможность генерировать фишинговые сообщения, трудноотличимые от созданных человеком [9, 10], а в сочетании с технологиями дипфейков, создавать реалистичные голосовые и видеоподделки для мошеннических схем [11].

Целью данной работы является экспериментальная демонстрация возможности использования публично доступной LLM для выполнения задач, связанных с penetration test (пентестом) и получением несанкционированного доступа в тестовую систему (специально уязвимая виртуальная машина Linux - Metasploitable2).

I. МЕТОДОЛОГИЯ ИССЛЕДОВАНИЯ

Для проверки гипотезы о том, что современные открытые LLM способны в автоматизированном режиме провести атаку был разработан и реализован лабораторный стенд, включающий атакующую и целевую системы. Атакующая система (Зоя) реализована через виртуальную машину (ВМ) с операционной системой (ОС) “Kali Linux”. Выбор обусловлен наличием в ней множества инструментов, предназначенных для пентеста. В качестве целевой (атакуемой) системы предполагается использование (Жанна) ВМ с ОС “Metasploitable2-Linux”. Такой вариант обусловлен наличием множества известных уязвимостей. На Рис. 1 представлено схематичное описание стенда.

Для реализации автоматизированного тестирования использованы Python-скрипты, обеспечивающие:

а) сканирование цели;

б) взаимодействие с LLM;

в) выполнение эксплоитов и проверка результатов их работы. Для этапа разведки будет применен инструмент Nmap.

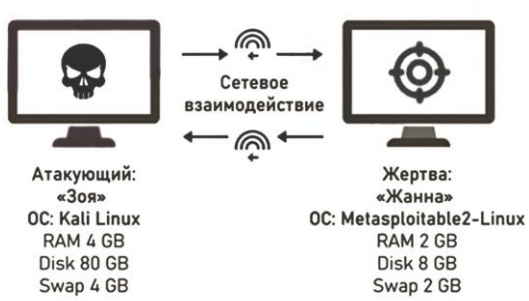


Рис. 1. Лабораторный стенд исследований

На Рис. 2 показан алгоритм работы скрипта.

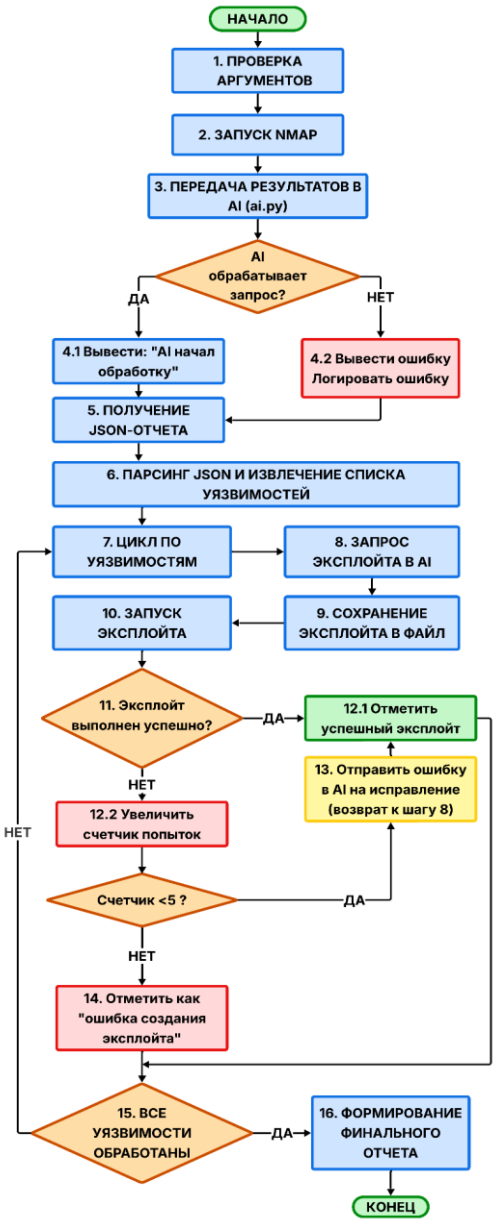


Рис. 2. Алгоритм работы скрипта

Важным фактором успеха использования LLM для совершения пентеста, следует считать факт, что попытка проникновения в систему выполняется через стандартные алгоритмы и использование определенных программ. Так, в рамках данной работы, применялся разработанный алгоритм:

а) потенциальный нарушитель запускает скрипт с указанием целевого IP-адреса;

б) скрипт запускает средство проведения аудита сети Nmap;

в) скрипт отправляет в LLM специальный формированный промпт и прикладывает отчет из Nmap;

г) LLM анализирует отчет о найденных сканером уязвимостях и формирует JSON-файл с описанием всех уязвимостей;

д) скрипт отправляет в LLM специальный промпт и прикладывает сформированный ранее JSON-файл,

е) LLM поэтапно генерирует эксплойты и отправляет их в скрипт;

ж) скрипт запускает поочередно эксплойты и проверяет было ли выполнено заявленное действие, если возникает ошибка отправляет в LLM ошибку с просьбой решить проблему, скрипт делает 5 попыток исправить эксплоит и идет дальше.

По итогу работы скрипта формируется отчет, включающий в себя перечень уязвимостей, результаты их эксплуатации, количество попыток и перечень возникших ошибок.

II. РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЙ

В ходе серии экспериментов был проведен многократный запуск скрипта автоматического пентеста, направленного на целевую систему “Жанна”. В результате экспериментов можно сделать вывод, что LLM практически в 100% случаев успешно анализировала вывод Nmap, вследствие чего верно были определены уязвимости. Это позволяло предлагать различные варианты эксплуатации обнаруженных уязвимостей. На этом этапе формировался JSON-файл (Рис. 3), на основе которого дальше LLM осуществляет попытки пентеста путем генерации и запуска эксплойтов.


```
{
  "target_ip": "192.168.13.35",
  "timestamp": "20260210_113027",
  "missing_packages": [],
  "vulnerabilities": [
    {
      "id": "vsftpd_234_backdoor",
      "порт": "21",
      "протокол": "tcp",
      "сервис": "vsftpd",
      "версия": "2.3.4",
      "уязвимость": "Backdoor Command",
      "cve": "CVE-2011-2523",
      "описание": "Версия vsftpd 2.3.4 содержит backdoor, активируемый через команду :)",
      "серьезность": "высокая",
      "ожидаемый_результат": "Получение обратной оболочки",
      "тип_эксплоита": "remote",
      "требуемые_библиотеки": [
        "socket",
        "sys",
        "os"
      ],
      "рекомендации": "Использовать Metasploit модуль 'exploit/unix/ftp/vsftpd_234_backdoor'"
    },
```


Рис. 3. Пример описания одной уязвимости от LLM

Стоит отметить, что несмотря на успешное определение уязвимостей и описание методов их

эксплуатации, LLM удалось лишь в 0,11% случаев использовать уязвимость vsFTPd 2.3.4 (2 из 1755). Для остальных 16-ти обнаруженных уязвимостей ни один эксплоит не оказался рабочим. Таким образом результаты эксперимента показали, что:

1. Нейросеть генерировала код с синтаксическими ошибками.

Например, отчёт показал ошибку выполнения эксплоита: NameError: name 'name' is not defined. Нейросеть при генерации кода использовала “name” вместо “__name__” в строчке “if __name__ == "__main__":”.

2. Нейросеть допускала синтаксические ошибки при работе с бинарными payload.

Например, отчёт показал ошибку выполнения эксплоита: SyntaxError: unterminated string literal.

3. Неспособность нейросети исправить ошибки, связанные с отсутствием требуемых зависимостей.

Например, отчёт показал ошибку выполнения эксплоита: ModuleNotFoundError: No module named 'mysql'. Несмотря на то, что скрипт позволял выполнить команды “pip install

…” нейросеть при получении подобной ошибки принимала решение переписать код, используя другие зависимости.

4. Некорректные эксплоит-реализации.

Даже в случае синтаксически правильного и исполняемого кода, эксплоит не выполнял поставленных задач.

5. Оборванные строковые литералы.

Например, отчёт показал ошибку выполнения эксплоита: payload += b"\x00\x00\x00\x00\x00\x

6. Нейросеть успешно структурировала атаку.

Нейросеть смогла планировать вектор атаки. В ходе эксперимента она успешно определила имеющиеся уязвимости и предложила правильные методики их эксплуатации.

Полученные результаты демонстрируют, что LLM способна выявлять уязвимости и формировать вектор атаки, а в редких случаях эксплуатировать известные уязвимости путем генерации работающих эксплоитов.

Анализ неудачных случаев использования сгенерированных эксплоитов продемонстрировал, что ключевыми проблемами использования современных языковых моделей для осуществления автоматизированных атак стали синтаксические ошибки и неполные payload-строки при генерации эксплоита. Это согласуется с выводами в исследованиях проблем использования кода, сгенерированного LLM [12, 13]. Синтаксические ошибки в коде являются галлюцинацией LLM, обусловленные недостаточным количеством примеров корректной конструкции в обучающих данных для конкретной модели. Неудача с Samba (Рис. 4) обусловлена генерацией чрезмерно длинных payload-строк, которые обрезались на этапе записи в файл. Это может быть связано с ограничением длины вывода модели. В то же время положительный опыт обнаружения и описания известных уязвимостей показывает двойственную природу LLM.

```
Уязвимость: usermap_script() Function Vulnerability
Порт: 139/445
Сервис: Samba
Статус: НЕУДАЧА
Попыток: N/A
Ошибка:   File "/home/kali/ai/temp_test_3_20260210_113027.py", line 7
    payload = b'\x00\x00\x00\x90\xFF\x53\x4D\x42\x72\x00\x00\x00\x00\x
```

Рис. 4. Неудача с Samba

Анализ случаев успешной эксплуатации уязвимости показал, что атаки с простой и формализованной логикой иногда могут выполняться LLM автоматически. В частности, уязвимость FTP-сервиса vsftpd 2.3.4 предполагает отправку специально сформированной строки, активирующей бэкдор. Такая уязвимость не требует установки зависимостей и генерации сложного кода. Остальные же уязвимости были подробно описаны, в том числе методы их эксплуатации. Следует отметить, что в случае исправления синтаксических ошибок в коде или установки требуемых зависимостей некоторые эксплойты оказались рабочими, так, например, была проэксплуатирована уязвимость «Metasploitable root shell» (bindshell).

ЗАКЛЮЧЕНИЕ

Результаты исследования позволяют сделать выводы о том, что языковая модель DeepSeek-R1 имеет значительные недостатки в качестве генерируемого кода, но тем не менее, в редких случаях способна совершить атаку без вмешательства человека. Можно сделать вывод о потенциале использования подобных моделей для совершения кибератак как минимум в качестве консультанта для начинающих хакеров. Это подтверждается и тем, что, внеся незначительные изменения в нерабочие эксплоиты, их можно было использовать для эксплуатации уязвимости. В дальнейшем исследовании планируется сфокусироваться на анализе генерируемого кода и изучении возможности его исправления в случае необходимости.

В результате можно сказать, что это исследование, как и другие [14, 15] показали, что современные языковые модели, подобные DeepSeek-R1, имеют ограниченные возможности в задачах автоматизированного пентестинга. Однако, даже при столь низкой эффективности можно уверенно говорить о возможности массового использовании LLM для автоматизированных атак на системы с документированными и простыми в реализации уязвимостями.

ЛИТЕРАТУРА

[1] Williams, S. Cybersecurity teams are preparing for a surge in global CVE vulnerabilities in 2026 // SecurityBrief Asia. — 2026. — Режим доступа: https://securitybrief.asia/story/cybersecurity-teams-brace-for-surge-in-global-cves-in-2026.

[2] Singer, B., Lucas, K., Adiga, L., Jain, M., Bauer, L., Sekar, V. (2025). Incalmo: An Autonomous LLM-Based System for Red Teaming in Multi-Component Networks. arXiv.org. DOI: 10.1007/s10207-024-00868-2

[3] Mirsky, Y., Demontis, A., Kotak, J., Shankar, R., Gelei, D., Yang, L., Zhang, X., Pintor, M., Lee, W., Elovici, Y., Biggio, B. (2023). The Threat of Offensive Artificial Intelligence to Organizations. Computers & Security.: https://www.sciencedirect.com/science/article/abs/pii/S0167404822003984

[4] Gupta, M., Gupta, M. (2025). Measuring AI Cybersecurity: A Comprehensive Framework for Understanding Offensive and Adversarial AI. AI and Ethics, 5, 883–910. DOI: 10.1007/s43681-024-00427-4

[5] Hu, Z., Beuran, R., Tan, Y. (2020). Automated Penetration Testing Using Deep Reinforcement Learning. In: Proceedings of the 2020 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW), Genoa, Italy, pp. 2–10. DOI: 10.1109/EuroSPW51379.2020.00010

[6] Nguyen, T. T., Reddi, V. J. (2021). Deep Reinforcement Learning for Cybersecurity. arXiv.org.: https://arxiv.org/abs/1906.05799

[7] CVE-2025-26210: Cross-Site Scripting (XSS) Vulnerability in DeepSeek AI Products (2025). Feedly.: https://feedly.com/cve/CVE-2025-26210

[8] Kelly, D. J., Long, J., Mylonas, A., Pitropakis, N., Buchanan, W. J. (2024). A Systematic Review of Research Using Artificial Intelligence for Open Source Intelligence (OSINT) Applications. International Journal of Information Security, 23, 2911–2938. DOI: 17487/RFC8377

[9] IBM aThink (2023). AI vs Human Deceit: Unravelling New-Age Phishing Tactics.: https://www.ibm.com/think/x-force/ai-vs-human-deceit-unravelling-new-age-phishing-tactics

[10] Инфобезопасность.ру (2025). ИИ вооружил киберпреступников — число атак в России выросло на треть.: https://infobezopasnost.ru/blog/news/ii-vooruzhil-kiberprestupnikov-chislo-atak-v-rossii-vyroslo-na-tret

[11] Как реклама с дипфейками в Instagram приводит к потере денег. (2025): https://www.kaspersky.ru/blog/scam-with-deepfakes-in-instagram-facebook-whatsapp/40221

[12] Zhang, Z., Wang, C., Wang, Y., Shi, E., Ma, Y., Zhong, W., Chen, J., Mao, M., Zheng, Z. (2025). LLM Hallucinations in Practical Code Generation: Phenomena, Mechanisms, and Mitigation. Proceedings of ISSTA 2025. DOI: 10.1145/3728894

[13] LLM they cannot cope with the search and exploitation of vulnerabilities // Infosecurity Magazine.: https://www.infosecurity-magazine.com/news/llms-fall-vulnerability-discovery

[14] Todd, D. (2025). DeepSeek and AI-Generated Malware Pose New Cybersecurity Threat. SecureWorld.: https://www.secureworld.io/industry-news/deepseek-ai-generated-malware

[15] Alger, J. (Managing Editor) (2025). DeepSeek Can Develop Malware: Cyber Experts Share the Risks. Security Magazine.: https://www.securitymagazine.com/articles/101470-deepseek-can-develop-malware-cyber-experts-are-sharing-the-risks

Информация об авторах

Азява Дмитрий Александрович, студент кафедры Защиты информации Новосибирского государственного технического университета, г. Новосибирск, Россия, e-mail: dimulatoraz@gmail.com, ORCID: 0009-0006-5107-9419.

Иллак Наталья Вячеславовна, студентка кафедры Защиты информации Новосибирского государственного технического университета, г. Новосибирск, Россия, e-mail: illak.nvs@gmail.com, ORCID: 0009-0003-1043-4611.

Киселев Антон Анатольевич, старший преподаватель кафедры Защиты информации Новосибирского государственного технического университета, г. Новосибирск, Россия, e-mail: anton.kiselev@corp.nstu.ru, ORCID: 0000-0002-0464-5039

Investigation of the possibility of conducting automated penetration testing with publicly available large language models

Dmitry Azyava, Natalia Illak, Anton Kiselev

Novosibirsk State Technical University, Novosibirsk, Russia

***Abstract* – The paper explores the possibility of using publicly available large language models to carry out cyber attacks by automated system penetration testing. The testing was conducted using the DeepSeek-R1 large language model. The experiment was conducted on a laboratory stand consisting of two virtual machines with operating systems: Kali Linux and Metasploitable2. The results showed the possibility of using a large language model for conducting a pentest, but with some limitations.**

***Keywords* – LLM, AI, offensive AI.**

REFERENCES

[1] Williams, S. Cybersecurity teams are preparing for a surge in global CVE vulnerabilities in 2026 // SecurityBrief Asia. — 2026. — Access mode: https://securitybrief.asia/story/cybersecurity-teams-brace-for-surge-in-global-cves-in-2026 (accessed: 01/13/2026).

[2] Singer, B., Lucas, K., Adiga, L., Jain, M., Bauer, L., Sekar, V. (2025). Incalmo: An Autonomous LLM-Based System for Red Teaming in Multi-Component Networks. arXiv.org. DOI: 10.1007/s10207-024-00868-2

[3] Mirsky, Y., Demontis, A., Kotak, J., Shankar, R., Gelei, D., Yang, L., Zhang, X., Pintor, M., Lee, W., Elovici, Y., Biggio, B. (2023). The Threat of Offensive Artificial Intelligence to Organizations. Computers & Security. Retrieved January 14, 2026, from https://www.sciencedirect.com/science/article/abs/pii/S0167404822003984

[4] Gupta, M., Gupta, M. (2025). Measuring AI Cybersecurity: A Comprehensive Framework for Understanding Offensive and Adversarial AI. AI and Ethics, 5, 883–910. DOI: 10.1007/s43681-024-00427-4

[5] Hu, Z., Beuran, R., Tan, Y. (2020). Automated Penetration Testing Using Deep Reinforcement Learning. In: Proceedings of the 2020 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW), Genoa, Italy, pp. 2–10. DOI: 10.1109/EuroSPW51379.2020.00010

[6] Nguyen, T. T., Reddi, V. J. (2021). Deep Reinforcement Learning for Cybersecurity. arXiv.org. Retrieved January 15, 2026, from https://arxiv.org/abs/1906.05799

[7] CVE-2025-26210: Cross-Site Scripting (XSS) Vulnerability in DeepSeek AI Products (2025). Feedly. Retrieved January 15, 2026,

from https://feedly.com/cve/CVE-25-2621
[8] Kelly, D. J., Long, J., Mylonas, A., Pitropakis, N., Buchanan, W. J. (2024). A Systematic Review of Research Using Artificial Intelligence for Open Source Intelligence (OSINT) Applications. International Journal of Information Security, 23, 2911–2938. DOI: 17487/RFC8377
[9] IBM aThink (2023). AI vs Human Deceit: Unravelling New-Age Phishing Tactics. Retrieved January 17, 2026, from https://www.ibm.com/think/x-force/ai-vs-human-deceit-unravelling-new-age-phishing-tactics
[10] Infobezopasnost′.ru (2025). II vooruzhil kiberprestupnikov — chislo atak v Rossii vyroslo na tret′. Retrieved January 17, 2026, from https://infobezopasnost.ru/blog/news/ii-vooruzhil-kiberprestupnikov-chislo-atak-v-rossii-vyroslo-na-tret
[11] Kak reklama s dipfejkami v Instagram privodit k potere deneg. (2025). Retrieved January 17, 2026, from https://www.kaspersky.ru/blog/scam-with-deepfakes-in-instagram-facebook-whatsapp/40221
[12] Zhang, Z., Wang, C., Wang, Y., Shi, E., Ma, Y., Zhong, W., Chen, J., Mao, M., Zheng, Z. (2025). LLM Hallucinations in Practical Code Generation: Phenomena, Mechanisms, and Mitigation. Proceedings of ISSTA 2025. DOI: 10.1145/3728894
[13] LLM they cannot cope with the search and exploitation of vulnerabilities // Infosecurity Magazine. — 2025. — Access mode: https://www.infosecurity-magazine.com/news/llms-fall-vulnerability-discovery / (date of access: 03/09/2026).
[14] Todd, D. (2025). DeepSeek and AI-Generated Malware Pose New Cybersecurity Threat. SecureWorld. Retrieved February 3, 2026, from https://www.secureworld.io/industry-news/deepseek-ai-generated-malware
[15] Alger, J. (Managing Editor) (2025). DeepSeek Can Develop Malware: Cyber Experts Share the Risks. Security Magazine. Retrieved February 8, 2026, from https://www.securitymagazine.com/articles/101470-deepseek-can-develop-malware-cyber-experts-are-sharing-the-risks

Information about the authors

Dmitry A. Azyava, student of the Information Security Department of Novosibirsk State Technical University, Novosibirsk, Russia, e-mail: dimulatoraz@gmail.com, ORCID: 0009-0006-5107-9419.

Natalia V. Illak, student of the Information Security Department of Novosibirsk State Technical University, Novosibirsk, Russia, e-mail: illak.nvs@gmail.com, ORCID: 0009-0003-1043-4611.

Anton A. Kiselev, Senior Lecturer at the Information Security Department of Novosibirsk State Technical University, Novosibirsk, Russia, e-mail: anton.kiselev@corp.nstu.ru, ORCID: 0000-0002-0464-5039.

Проектирование системы обнаружения сетевых атак на базе решений с открытым исходным кодом Snort и ELK Stack

В.Д. Зимнюкова, А.С. Ершов

Саратовский государственный технический университет имени Гагарина Ю.А., Саратов, Россия

Аннотация **– В статье рассматривается процесс проектирования и практической реализации интегрированной системы обнаружения сетевых атак на основе решений с открытым исходным кодом – Snort и ELK Stack. Проводится анализ современных подходов к мониторингу информационной безопасности, исследуются требования нормативных документов ФСТЭК России и российского законодательства к системам обнаружения вторжений и управления событиями безопасности. Описывается архитектура лабораторного стенда, включающая четыре виртуальные машины с распределением ролей: сенсор Snort IDS, сервер ELK Stack, целевой сервер-жертву и машина атакующего. Демонстрируется процесс настройки пересылки логов, разработки правил обнаружения и создания визуализаций в Kibana. Представлены результаты тестирования системы при проведении различных типов атак: сканирование сети, DoS-атаки, веб-атаки, подбор учетных данных и использование сканера уязвимостей. Анализ эффективности разработанной системы подтверждает высокую точность обнаружения и практическую применимость предложенного подхода для организаций с ограниченными ресурсами.**

Ключевые слова **– обнаружение сетевых атак, Snort, ELK Stack, SIEM, IDS, мониторинг информационной безопасности, кибератака, система мониторинга ИБ, система обнаружения вторжений.**

ВВЕДЕНИЕ

В условиях цифровой трансформации и повсеместного распространения информационных технологий вопросы обеспечения информационной безопасности приобретают критическую важность. Современные организации сталкиваются с беспрецедентным ростом количества и сложности сетевых атак, которые становятся все более изощренными и целенаправленными. Особую тревогу вызывает эволюция методов киберпреступников — от массовых автоматизированных атак к целевым сложносоставным кампаниям, использующим социальную инженерию и продвинутые техники уклонения от обнаружения. Традиционные системы защиты периметра оказываются недостаточно эффективными против современных угроз, требующих комплексного подхода к мониторингу безопасности.

В этой связи особую актуальность приобретают системы обнаружения вторжений (Intrusion Detection System, IDS), способные выявлять аномальную активность внутри сетевой инфраструктуры, и системы управления событиями безопасности (Security Information and Event Management, SIEM), обеспечивающие централизованный сбор, корреляцию и анализ событий безопасности. Однако их разрозненное использование не позволяет в полной мере реализовать потенциал современных технологий защиты.

I. ПОСТАНОВКА ЗАДАЧИ

Целью данной работы является проектирование и практическая реализация интегрированной системы обнаружения и визуализации сетевых атак на базе решений с открытым исходным кодом. Такой подход позволяет не только продемонстрировать эффективность комплексного мониторинга безопасности, но и предложить экономически эффективное решение для организаций различного масштаба.

II. СИСТЕМЫ ОБНАРУЖЕНИЯ ВТОРЖЕНИЙ И УПРАВЛЕНИЯ СОБЫТИЯМИ БЕЗОПАСНОСТИ

Системы обнаружения вторжений функционируют на основе анализа сетевого трафика в режиме, близком к реальному времени [1]. Основополагающим принципом является пассивный мониторинг передаваемых данных без воздействия на сетевые потоки. Система осуществляет захват пакетов, их декодирование и последующий анализ с применением методов выявления аномалий. Сигнатурный анализ позволяет идентифицировать известные угрозы путем сравнения с эталонными шаблонами атак, в то время как поведенческие методы направлены на обнаружение

отклонений от базового профиля нормальной сетевой активности [1].

Традиционные системы обнаружения вторжений обладают рядом принципиальных ограничений, сужающих область их эффективного применения в современных условиях. Наиболее существенным недостатком является контекстная изоляция анализа, при которой IDS обрабатывает сетевые события обособленно, без учета данных от других источников безопасности. Значительной проблемой остается высокий уровень ложных срабатываний, требующий постоянного вмешательства аналитиков для верификации инцидентов и тонкой настройки правил детектирования.

Эволюция подходов к мониторингу информационной безопасности демонстрирует закономерный переход от изолированных систем обнаружения вторжений к комплексным платформам управления событиями безопасности. SIEM-системы представляют собой платформы для централизованного сбора, хранения, анализа и корреляции событий безопасности из разнородных источников [2]. Ключевой функцией современных SIEM-решений является агрегация данных безопасности из различных систем. Сбор и нормализация логов позволяют привести разноформатные данные к единому стандарту, что существенно облегчает их последующий анализ.

Интеграция систем обнаружения вторжений и платформ управления событиями безопасности создает взаимоусиливающий эффект, значительно повышающий эффективность защиты информационной инфраструктуры. Основное преимущество интеграции заключается в возможности комплексного анализа данных безопасности из различных источников, что позволяет выявлять сложные целевые атаки, остающиеся незамеченными при раздельном использовании этих систем. Архитектура системы с интеграцией SIEM и IDS изображена на Рис. 1.

Рис. 1. Архитектура системы с интеграцией SIEM и IDS

III. ТРЕБОВАНИЯ К МОНИТОРИНГУ ИНФОРМАЦИОННОЙ БЕЗОПАСНОСТИ В РОССИИ

Основополагающим документом, регламентирующим процессы мониторинга, является ГОСТ Р 59547-2021 «Защита информации. Мониторинг информационной безопасности». Данный стандарт устанавливает требования к системам мониторинга информационной безопасности, включая сбор, нормализацию, анализ и корреляцию событий безопасности, а также определяет порядок их функционирования в автоматизированных системах различного уровня [3].

Приказ ФСТЭК России от 18 февраля 2013 г. № 21 определяет состав и содержание организационных и технических мер по обеспечению безопасности персональных данных при их обработке в информационных системах. Данный документ включает требования по реализации мер по обнаружению (предотвращению) вторжений, а также мер по мониторинг (просмотр, анализ) результатов регистрации событий безопасности и реагирование на них [4].

С 1 марта 2026 вступил в силу приказ ФТЭК 117 от 11 апреля 2025 г. «Об утверждении требований о защите информации, содержащейся в государственных информационных системах, иных информационных системах государственных органов, государственных унитарных предприятий, государственных учреждений». В данном приказе указано, что для достижения целей защиты информации оператором (обладателем информации) должны проводиться мероприятия по осуществлению мониторинга информационной безопасности в соответствии с разделами 4 и 5 ГОСТ Р 59547-2021 [5,6].

Совокупность рассмотренных документов формирует комплексную систему требований, обеспечивающих создание эффективных систем обнаружения и противодействия кибератакам. Соответствие выбранных в данном проекте решений требованиям ФСТЭК обеспечивается за счет использования Snort (система предотвращения и обнаружения вторжений), поддерживающего необходимые функции обнаружения и документирования атак в соответствии с мерой 8.7 Приказа № 21, и ELK Stack (набор инструментов для обработки логов, полученных системой обнаружения вторжений) предоставляющего требуемые возможности сбора, хранения, нормализации, анализа и визуализации событий безопасности в соответствии с мерой 8.5 и требованиями ГОСТ Р 59547-2021. Таким образом, предлагаемое решение не только демонстрирует техническую эффективность, но и учитывает требования российского законодательства в области мониторинга информационной безопасности.

IV. ОБОСНОВАНИЕ ВЫБОРА ТЕХНОЛОГИЙ ДЛЯ ПРОЕКТА

Выбор технологий для практической реализации системы обнаружения сетевых атак осуществлялся на основе комплексного анализа технических характеристик, экономической целесообразности и образовательной ценности решений. В качестве базовых технологий были выбраны Snort для функций обнаружения вторжений и ELK Stack для задач сбора, обработки и визуализации событий безопасности.

Выбор технологии Snort обусловлен ее технической зрелостью, подтвержденной более чем 20-летней историей развития, что обеспечивает стабильность работы и предсказуемость поведения системы. Архитектура Snort, основанная на модульном принципе, позволяет глубоко изучить механизмы работы современных систем обнаружения атак, начиная от захвата сетевого трафика и заканчивая применением сложных правил корреляции [7]. Открытая модель разработки правил обнаружения обеспечивает неограниченные возможности для создания специализированных сигнатур, адаптированных под конкретные исследовательские задачи.

Выбор технологии ELK Stack основан на ее исключительной гибкости и мощных возможностях аналитики. Модульная архитектура, состоящая из набора инструментов Elasticsearch, Logstash и Kibana, предоставляет полный цикл обработки данных безопасности – от приема и нормализации логов до их индексирования, хранения и интерактивной визуализации [8]. Компонент Elasticsearch обеспечивает высокопроизводительное полнотекстовое индексирование и поиск по большим объемам данных, программа Logstash предлагает богатый набор фильтров для преобразования данных, а Kibana предоставляет развитые средства построения дашбордов [8].

V. ПРОЕКТИРОВАНИЕ ЛАБОРАТОРНОГО СТЕНДА

Архитектура лабораторного стенда включает четыре виртуальные машины с четким распределением ролей и специализированным подбором операционных систем. Сервер ELK Stack выполняет функции центрального узла сбора и анализа событий безопасности на базе Ubuntu Server 22.04 LTS. Сенсор Snort IDS осуществляет мониторинг сетевого трафика, выявляя потенциально вредоносную активность, и также функционирует под управлением Ubuntu Server 22.04 LTS. Целевой сервер на базе уязвимой виртуальной машины Metasploitable 2, а машина атакующего использует операционную систему Kali Linux 2025.3 для генерации тестового вредоносного трафика.

Организация сетевой инфраструктуры построена на принципах сегментации и контролируемого взаимодействия между компонентами. Все виртуальные машины объединены в изолированную сеть с адресным пространством 192.168.100.0/24, что исключает непреднамеренное воздействие на внешние системы и обеспечивает воспроизводимость экспериментов. Сенсор Snort IDS с адресом 192.168.100.20 выполняет пассивный мониторинг всего сетевого трафика, сервер ELK Stack на адресе 192.168.100.10 принимает события через syslog-протокол, целевой сервер с адресом 192.168.100.30 представляет собой уязвимую систему, а машина атакующего с адресом 192.168.100.40 генерирует тестовые атаки. Схема виртуального стенда представлена на Рис. 2.

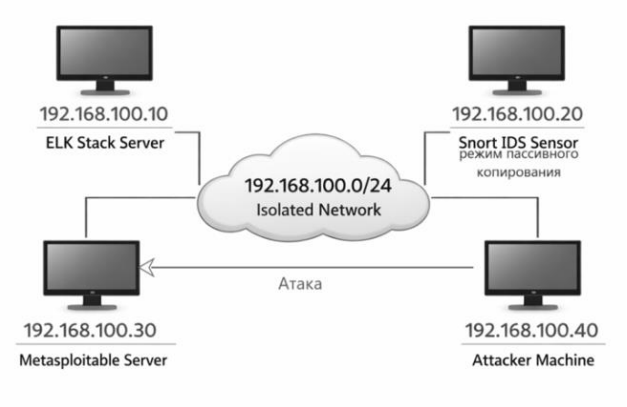


Рис. 2. Схема виртуального стенда

Критически важным аспектом являлась настройка режима promiscuous mode для сетевого интерфейса сенсора Snort IDS, позволяющего принимать и анализировать весь трафик, проходящий через сетевой сегмент, независимо от адреса назначения.

Для эффективного выявления различных типов атак в виртуальном стенде был разработан комплекс кастомных правил для сенсора Snort. При создании правил была изучена официальная документация и применялись различные опции: анализ флагов TCP, поиск специфического содержимого в пакетах, пороговые значения для обнаружения аномальной активности.

В ходе работы были разработаны правила для обнаружения следующих типов атак: подбор учётных данных (SSH, FTP, RDP), сканирование сети и портов (ICMP-запросы, SYN-, FIN- и XMAS-сканирование), веб-атаки (Directory Traversal, SQL-инъекции, XSS, Local File Inclusion, Remote File Inclusion, попытки загрузки веб-шеллов), атаки типа «отказ в обслуживании» (SYN-флуд, UDP-флуд), а также подозрительная активность. Для каждого правила подобраны пороговые значения, исключающие ложные срабатывания на фоновом трафике виртуальной сети. Разберем синтаксис правил на примере правила распознавания SYN-сканирования: alert tcp $EXTERNAL_NET any -> $HOME_NET any (msg:"SYN

Scan Detection"; flags:S; detection_filter:track by_src, count 10, seconds 10; sid:1000006; rev:1;).

alert – сгенерировать алерт.

tcp – только TCP-пакеты.

$EXTERNAL_NET any – любой внешний IP, любой порт источника.

-> – направление от источника к получателю.

$HOME_NET any – защищаемая сеть (любой IP), любой порт назначения.

msg – текст алерта.

flags:S – установлен только флаг SYN.

detection_filter: track by_src, count 10, seconds 10 – сработает, когда от одного IP-источника за 10 секунд будет 10 или более таких пакетов.

sid, rev – идентификатор и версия.

VI. ТЕСТИРОВАНИЕ СИСТЕМЫ

Методика тестирования разработанной системы была основана на комплексном подходе, включающем проведение контролируемых атак различных типов с последующей оценкой эффективности их обнаружения. План проведения тестовых атак предусматривал последовательное выполнение сценариев, имитирующих реальные угрозы информационной безопасности.

A. Сканирование сети и портов

Первый этап тестирования включал сканирование сети и портов с использованием утилиты nmap.

nmap -sU 192.168.100.30 -p 53,161,123,137,138,139,500,514 – UPD-сканирование.

nmap -sS 192.168.100.30 – TCP-сканирование.

После выполнения атак с Kali Linux на Metasploitable дашборды Kibana показали обнаружение SNMP и DNS сканирования с атакующей машины 192.168.100.40 (Рис. 3). В логах видны запросы к UDP-порту 161 (SNMP), TCP-порту 161 (SNMP), TCP-порту 705 (SNMP AgentX) и UDP-порту 53 (DNS). Это соответствует выполненным командам nmap -sU и nmap -sS.

Рис. 3. Изменения дашбордов после проведения сканирования сети и портов

B. Dos-атака

Были проведены тестовые DoS-атаки с использованием hping3 на порт 80 с флудом SYN-пакетов.

sudo hping3 -S –flood -p 80 192.168.100.30

После проведения Dos-атак временной график зафиксировал всплеск алертов в период атаки, доля TCP-трафика возросла до 90,7%, появились алерты категорий "BAD-TRAFFIC tcp port 0 traffic" и "MISC Source Port 0/53 to +1024", характерные для SYN-флуд атак (Рис. 4).

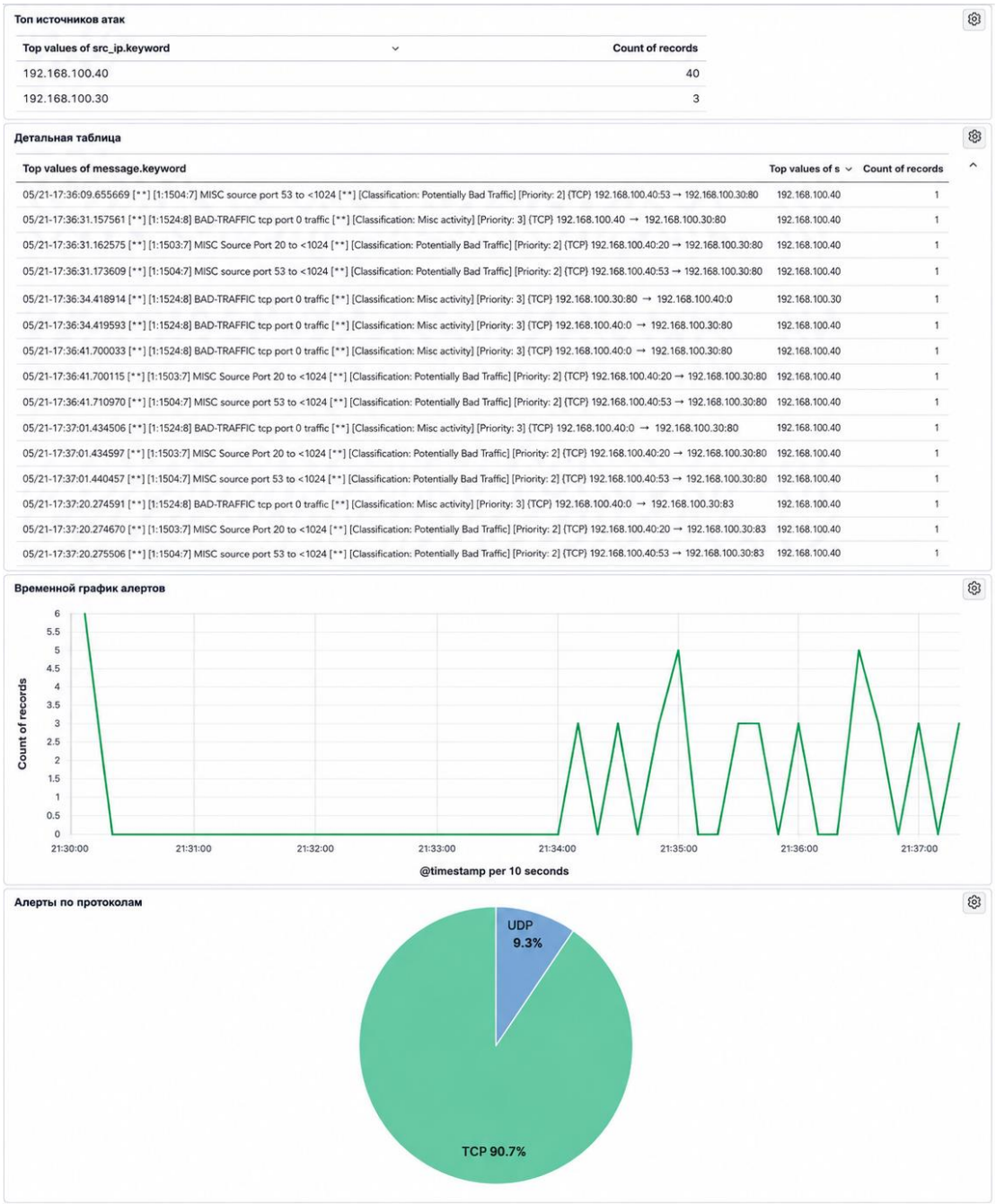


Рис. 4. Изменения дашбордов после проведения Dos-атаки

C. Веб-атаки

Были проведены веб-атаки, включая попытки доступа к /etc/passwd, /etc/shadow и XSS-атаку через параметр search.php.

curl “http://192.168.100.30/../../../etc/passwd” – Path Traversal

curl “http://192.168.100.30/../../../etc/shadow” – Path Traversal

curl “http://192.168.100.30/search.php?q=<script>alert(‘XSS’)</script>” – внедрение тега <script>alert('xss')</script> в параметр q запроса.

Дашборды продемонсрировали следующие изменения: сенсор Snort успешно выявил все три типа веб-атак, количество записей увеличилось, на графике появились новые пики в периоды выполнения атак (Рис. 5).

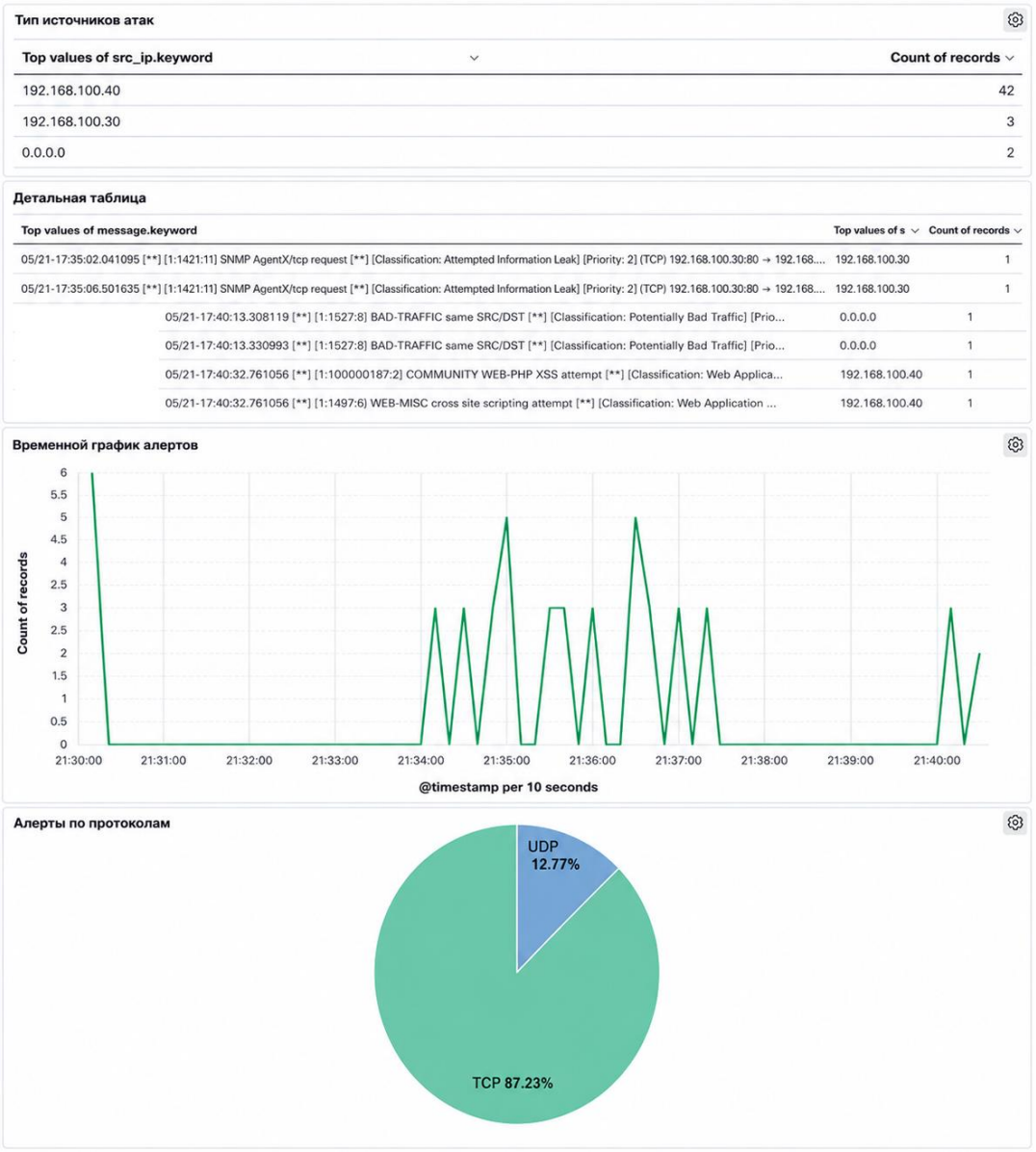


Рис. 5. Изменения дашбордов после проведения Веб-атак

D. Использование сканера уязвимостей

Использовали сканер уязвимостей dirb для сканирования веб-директорий.

dirb http://192.168.100.30

Дашборды отражают значительные изменения: временной график показывает несколько пиковых скачков, появились алерты от целевого сервера, протокольное распределение является полностью TCP-ориентированным.

На Рис. 6 показаны зафиксированные системой события, которые произошли из-за проведения этой атаки.

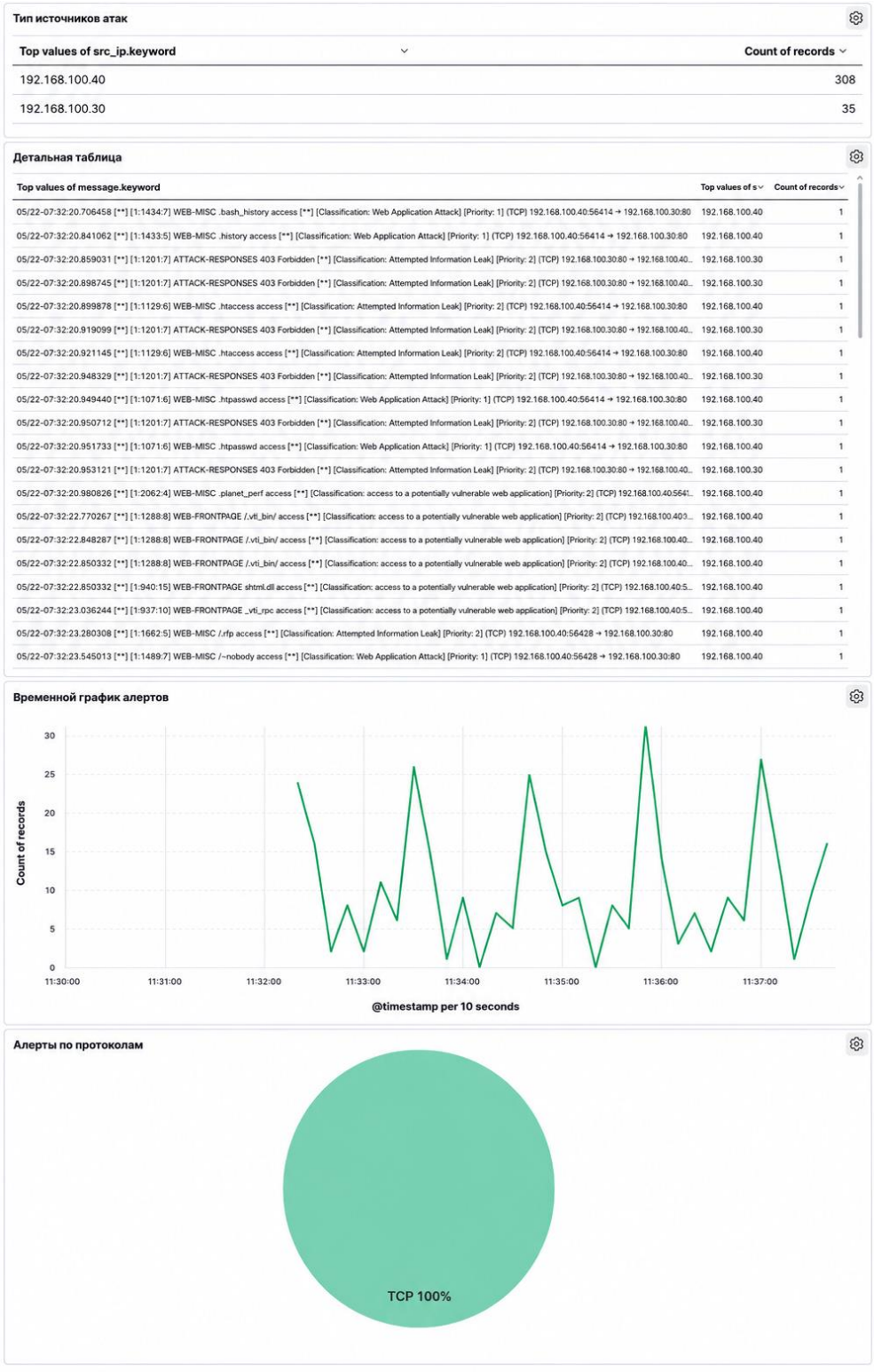


Рис. 6. Изменения дашбордов после проведения атаки с использованием сканера уязвимостей

Е. Атаки на аутентификацию

Были проведены тестовые атаки на аутентификацию методом грубой силы на SSH и FTP сервисы с использованием инструментов для подбора паролей методом грубой силы – Hydra и Medusa.

hydra -l annonymous -P test_pass.txt -t 2 ftp://192.168.100.30

medusa -h 192.168.100.30 -u root -P test_pass.txt -M ssh

Дашборд демонстрирует изменения после проведения атак: появились алерты "INFO FTP Bad login" с источником 192.168.100.30, показывающие направление трафика от сервера к клиенту, что подтверждает фиксацию системой реакции сервисов на атаки (Рис. 7).

Детальная таблица

Top values of message.keyword	Top values of src_ip.keyword	Count of records
12/07-10:23:29.579984 [**] [1:491:8] INFO FTP Bad login [**] [Classification: Potentially Bad Traffic] [Priority: 2] {TCP} 192.168.100.30:21 → 192.168.100.40:34616	192.168.100.30	1
12/07-10:23:29.579984 [**] [1:491:8] INFO FTP Bad login [**] [Classification: Potentially Bad Traffic] [Priority: 2] {TCP} 192.168.100.30:21 → 192.168.100.40:34628	192.168.100.30	1
12/07-10:23:33.236749 [**] [1:491:8] INFO FTP Bad login [**] [Classification: Potentially Bad Traffic] [Priority: 2] {TCP} 192.168.100.30:21 → 192.168.100.40:34616	192.168.100.30	1
12/07-10:23:33.356766 [**] [1:491:8] INFO FTP Bad login [**] [Classification: Potentially Bad Traffic] [Priority: 2] {TCP} 192.168.100.30:21 → 192.168.100.40:34628	192.168.100.30	1
12/07-10:23:36.526254 [**] [1:491:8] INFO FTP Bad login [**] [Classification: Potentially Bad Traffic] [Priority: 2] {TCP} 192.168.100.30:21 → 192.168.100.40:34616	192.168.100.30	1
12/07-10:23:36.653649 [**] [1:491:8] INFO FTP Bad login [**] [Classification: Potentially Bad Traffic] [Priority: 2] {TCP} 192.168.100.30:21 → 192.168.100.40:34628	192.168.100.30	1

Рис. 7. Изменения дашбордов после проведения атаки на аутентификацию

VII. АНАЛИЗ РЕЗУЛЬТАТОВ И ВЫВОДЫ

Анализ результатов тестирования продемонстрировал высокую эффективность обнаружения различных типов атак разработанной системой. Все запланированные тестовые атаки были успешно обнаружены с минимальным временем задержки между началом атаки и появлением соответствующего оповещения в платформе визуализации данных Kibana. Временные характеристики системы варьировались в зависимости от типа атаки – простые атаки сканирования обнаруживались практически мгновенно, в то время как сложные многоэтапные атаки требовали накопления статистики для срабатывания пороговых правил.

Точность срабатывания правил оказалась исключительно высокой – все кастомные правила, разработанные в рамках проекта, продемонстрировали корректную работу.

ЗАКЛЮЧЕНИЕ

Практическая значимость проведенного исследования заключается в создании работоспособного прототипа системы мониторинга информационной безопасности, который может быть развернут в организациях с ограниченными финансовыми ресурсами без потери функциональности. Предложенная архитектура легко масштабируется за счет добавления дополнительных сенсоров Snort в различные сегменты сети, а гибкость набора инструментов ELK Stack позволяет подключать новые источники событий безопасности без существенной модификации существующей конфигурации.

ЛИТЕРАТУРА

[1] Академия Selectel – IPS/IDS — системы обнаружения и предотвращения вторжений [Электронный ресурс] – Режим доступа: https://selectel.ru/blog/ips-and-ids/, свободный (дата обращения: 13.02.2026).

[2] Академия Selectel – Что такое SIEM [Электронный ресурс] – Режим доступа: https://selectel.ru/blog/what-is-siem/, свободный (дата обращения: 13.02.2026).

[3] ГОСТ Р 59547-2021 «Защита информации. Мониторинг информационной безопасности».

[4] Приказ ФСТЭК России от 18 февраля 2013 г. №21 «Об утверждении состава и содержания организационных и технических мер по обеспечению безопасности персональных данных при их обработке в информационных системах персональных данных».

[5] 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».

[6] Информационное письмо ФСТЭК России «Об утверждении требований к системам обнаружения вторжений».

[7] Snort – Network Intrusion Detection & Prevention System [Электронный ресурс] – Режим доступа: https://snort.org, свободный (дата обращения: 10.02.2026). Kaspersky SecureList Статистика [Электронный ресурс] – Режим доступа: https://statistics.securelist.com/ru/intrusion-detection-scan/day, свободный (дата обращения: 13.02.2026).

[8] ELK Stack – Elastic Stack: (ELK) Elasticsearch, Kibana & Logstash | Elastic [Электронный ресурс] – Режим доступа: https://www.elastic.co/elastic-stack, свободный (дата обращения: 10.02.2026).

Информация об авторах

Зимнюкова Вероника Дмитриевна, студентка четвертого курса, учебной группы б2-ИФБС-41, направления 10.03.01 «Информационная безопасность», Саратовского государственного

технического университета имени Гагарина Ю.А., г. Саратов, Россия, veronikazimnukova@gmail.com

Ершов Алексей Сергеевич, доцент кафедры ИБС, к.т.н., Саратовского государственного технического университета, имени Гагарина Ю.А., г. Саратов, Россия, ershovas@sstu.ru.

Design of the network attacks detection system based on Snort and ELK Stack

Veronika Zimnyukova, Alexey Ershov

Yuri Gagarin Saratov State Technical University, Saratov, Russia

***Abstract* – The article discusses the process of designing and practical implementation of an integrated network attack detection system based on open-source solutions – Snort and ELK Stack. The analysis of modern approaches to monitoring of information security is carried out, the requirements of regulatory documents of the FSTEC of Russia and Russian legislation to intrusion detection systems and management of security events are investigated. The architecture of a laboratory stand is described, including four virtual machines with role distribution: Snort IDS sensor, ELK Stack server, target server-victim and machine of the attacker. The article demonstrates the process of configuring log forwarding, developing detection rules, and creating visualizations in Kibana. The article presents the results of testing the system during various types of attacks, including network scanning, DoS attacks, web attacks, credential harvesting, and vulnerability scanning. The analysis of the system's effectiveness confirms its high accuracy.**

***Keywords* – network attack detection, Snort, ELK Stack, SIEM, IDS, information security monitoring, cyber-attack, information security monitoring system, intrusion detection system.**

REFERENCES

[1] Akademiya Selectel – IPS/IDS — sistemy obnaruzheniya i predotvrashcheniya vtorzhenij [Elektronnyj resurs] – Rezhim dostupa: https://selectel.ru/blog/ips-and-ids/, svobodnyj (data obrashcheniya: 13.02.2026).

[2] Akademiya Selectel – Chto takoe SIEM [Elektronnyj resurs] – Rezhim dostupa: https://selectel.ru/blog/what-is-siem/, svobodnyj (data obrashcheniya: 13.02.2026).

[3] GOST R 59547-2021 «Zashchita informacii. Monitoring informacionnoj bezopasnosti».

[4] Prikaz FSTEK Rossii ot 18 fevralya 2013 g. №21 «Ob utverzhdenii sostava i soderzhaniya organizacionnyh i tekhnicheskih mer po obespecheniyu bezopasnosti personal'nyh dannyh pri ih obrabotke v informacionnyh sistemah personal'nyh dannyh».

[5] Prikaz FSTEK Rossii ot 11 aprelya 2025 g. №117 «Ob utverzhdenii trebovanij o zashchite informacii, soderzhashchejsya v gosudarstvennyh informacionnyh sistemah, inyh informacionnyh sistemah gosudarstvennyh organov, gosudarstvennyh unitarnyh predpriyatij, gosudarstvennyh uchrezhdenij».

[6] Informacionnoe pis'mo FSTEK Rossii «Ob utverzhdenii trebovanij k sistemam obnaruzheniya vtorzhenij».

[7] Snort – Network Intrusion Detection & Prevention System [Elektronnyj resurs] – Rezhim dostupa: https://snort.org, svobodnyj (data obrashcheniya: 10.02.2026). Kaspersky SecureList Statistika [Elektronnyj resurs] – Rezhim dostupa: https://statistics.securelist.com/ru/intrusion-detection-scan/day, svobodnyj (data obrashcheniya: 13.02.2026).

[8] ELK Stack – Elastic Stack: (ELK) Elasticsearch, Kibana & Logstash | Elastic [Elektronnyj resurs] – Rezhim dostupa: https://www.elastic.co/elastic-stack, svobodnyj (data obrashcheniya: 10.02.2026).

Information about the authors

Zimnyukova Veronika Dmitrievna, fourth-year student, study group b2-IFBS-41, direction 10.03.01 "Information Security", Yuri Gagarin State Technical University of Saratov, Saratov, Russia, veronikazimnukova@gmail.com

Ershov Alexey Sergeevich, Associate professor of the Department of Information Security, Candidate of Technical Sciences, Yuri Gagarin State Technical University of Saratov, Saratov, Russia, ershovas@sstu.ru.

Разработка модуля автоматизированного мониторинга доступности серверов для малого предприятия

А.А. Писаренко

Донской государственный технический университет, Ростов-на-Дону, Россия

Аннотация **– рассматривается проблема обеспечения непрерывного контроля состояния серверной инфраструктуры на малых предприятиях. Предложен оригинальный программный модуль мониторинга доступности серверов, реализованный на языке Python с применением асинхронных библиотек asyncio и aiohttp. Выполнен сравнительный анализ существующих решений – Zabbix, Prometheus, Nagios. Описана архитектура разработанного модуля, его RESTful API и веб-интерфейс. Проведено тестирование на десяти серверах с использованием протоколов ICMP, HTTP и TCP. Получены результаты, свидетельствующие о высокой производительности и низком пороге вхождения при внедрении.**

Ключевые слова **– мониторинг серверов, малое предприятие, Python, asyncio, ICMP, HTTP, TCP, REST API, веб-интерфейс**

ВВЕДЕНИЕ

Инфраструктура информационных технологий является неотъемлемым элементом операционной деятельности современных предприятий независимо от их масштаба. Для организаций малого бизнеса надёжность работы серверного парка напрямую определяет непрерывность бизнес-процессов: отказ почтового сервера, веб-приложения или корпоративной базы данных даже на несколько минут может повлечь финансовые потери и репутационный ущерб [1].

Тем не менее, малые предприятия, как правило, располагают ограниченными ресурсами для внедрения и сопровождения комплексных систем мониторинга. Существующие решения корпоративного класса – Zabbix, Nagios, Prometheus – обладают широкими функциональными возможностями, однако их развёртывание и дальнейшая поддержка требуют значительных временных и финансовых затрат, а также высокой квалификации персонала [2]. По данным исследования аналитической компании Gartner (2023), около 67% малых предприятий в России отказываются от внедрения профессиональных средств мониторинга именно по причине сложности их первоначальной настройки [3].

Таким образом, актуальной является задача разработки лёгкого, простого в установке и эксплуатации модуля мониторинга, ориентированного на нужды малых предприятий и не требующего специализированной инфраструктуры. Целью настоящей работы является проектирование, программная реализация и тестирование такого модуля на основе современного стека технологий Python [4].

I. ПОСТАНОВКА ЗАДАЧИ

Целью работы является разработка и экспериментальное подтверждение эффективности лёгкого программного модуля мониторинга доступности серверов, предназначенного для малых предприятий с ограниченными ИТ-ресурсами. Для достижения цели необходимо провести сравнительный анализ существующих систем мониторинга (Zabbix, Prometheus, Nagios) с выделением их недостатков для малого бизнеса, спроектировать и реализовать модуль на языке Python с использованием асинхронного подхода и поддержкой протоколов ICMP, HTTP и TCP, обеспечить простой веб-интерфейс и REST API, а затем выполнить экспериментальную проверку на стенде из десяти виртуальных серверов в течение 72 часов. Ключевыми оцениваемыми параметрами являются время отклика, нагрузка на центральный процессор и оперативную память, а также время уведомления о сбое. Результаты сравниваются с корпоративными аналогами по критериям простоты установки, стоимости внедрения и достаточности функциональности для малого предприятия.

II. ОБЗОР СУЩЕСТВУЮЩИХ РЕШЕНИЙ

На рынке средств мониторинга IT-инфраструктуры выделяют три наиболее распространённых решения открытого класса: Zabbix, Prometheus и Nagios. Каждое из них имеет характерные архитектурные особенности и целевую аудиторию.

В Табл. I приведено сравнение рассмотренных решений с предлагаемым модулем по ключевым критериям, значимым для малого предприятия.

ТАБЛИЦА I
СРАВНЕНИЕ ХАРАКТЕРИСТИК СИСТЕМ МОНИТОРИНГА

Критерий	Zabbix	Promet-heus	Nagios	Предл. модуль	Вес (%)
Простота установки	3/5	4/5	2/5	5/5	20
Поддержка протоколов	5/5	4/5	5/5	4/5	20
Веб-интерфейс	4/5	3/5	3/5	5/5	15
Масштабируемость	5/5	5/5	4/5	3/5	15
Стоимость внедрения	3/5	5/5	3/5	5/5	20
Документация	4/5	4/5	3/5	4/5	10
Итоговый балл	3,95	4,20	3,30	4,55	100

Zabbix – масштабируемая корпоративная система мониторинга с открытым исходным кодом, разработанная компанией Zabbix SIA. Её архитектура основана на клиент-серверной модели с использованием агентов, устанавливаемых на контролируемые узлы. Система поддерживает активные и пассивные проверки, автоматическое обнаружение ресурсов (auto-discovery), гибкую систему триггеров и многоуровневые уведомления. Веб-интерфейс Zabbix реализован на PHP и отличается высокой информативностью, однако обладает относительно высоким порогом освоения [5]. Для малого предприятия с 10–15 серверами Zabbix может быть избыточен: его установка занимает от 4 до 8 часов, а начальная конфигурация требует глубоких знаний в области сетевого администрирования.

Prometheus – система мониторинга и база данных временных рядов, разработанная в рамках экосистемы Cloud Native Computing Foundation. В отличие от Zabbix, Prometheus использует модель pull: сервер самостоятельно опрашивает конечные точки (exporters) по протоколу HTTP. Ключевыми достоинствами являются нативная интеграция с Kubernetes, мощный язык запросов PromQL и широкая экосистема экспортёров. Однако отсутствие встроенного инструмента визуализации – Prometheus обычно используется совместно с Grafana [6] – усложняет развёртывание в условиях ограниченных ресурсов малого бизнеса [7].

Nagios является одним из старейших инструментов мониторинга с открытым кодом. Его архитектура строится на системе плагинов, что обеспечивает высокую гибкость, однако одновременно создаёт значительную сложность конфигурации. Интерфейс Nagios Core морально устарел, а большинство современных возможностей доступны лишь в коммерческой редакции Nagios XI. Несмотря на широкую распространённость, Nagios уступает конкурентам по удобству начальной настройки [8].

III. АРХИТЕКТУРА ПРЕДЛАГАЕМОГО МОДУЛЯ

При проектировании модуля было важно учесть следующее практическое требование: система должна запускаться без многочасовой настройки на любом Linux-хосте (включая дешёвый VPS). Именно это требование определило принцип минимально достаточной функциональности [9]. Архитектурно модуль делится на три уровня: сбор данных, хранение с обработкой и представление, хотя на практике граница между первыми двумя размыта, планировщик одновременно инициирует проверки и записывает результаты.

Бэкенд реализован на Python 3.11. Центральным проектным решением стал отказ от многопоточности в пользу асинхронной модели на базе asyncio [10], это позволяет параллельно опрашивать десятки узлов в рамках одного процесса ОС. HTTP-запросы обрабатывает aiohttp, ICMP-пакеты – icmplib (для работы с raw-сокетами процессу назначается capability CAP_NET_RAW). Расписание проверок ведёт APScheduler: в отличие от простого asyncio.sleep в цикле, он корректно выдерживает интервал даже при кратковременных зависаниях отдельных хостов. REST API построен на FastAPI [11], история состояний пишется в SQLite через aiosqlite – файловая БД здесь достаточна, а развёртывание не требует отдельного сервера СУБД [12].

IV. ОПИСАНИЕ API

API включает восемь конечных точек, которые покрывают все сценарии эксплуатации и приведены в Табл. II.

ТАБЛИЦА II
ПЕРЕЧЕНЬ КОНЕЧНЫХ ТОЧЕК REST API

Метод	Путь	Описание
GET	/api/servers	список всех серверов
POST	/api/servers	добавить сервер
DELETE	/api/servers/{id}	удалить сервер
GET	/api/servers/{id}/status	текущий статус
GET	/api/servers/{id}/history	история проверок
POST	/api/servers/{id}/check	внеплановая проверка
GET	/api/alerts	список активных алертов
POST	/api/settings	конфигурация интервалов

Аутентификация через Bearer JWT. Срок жизни токена выставлен коротким (15 минут) с поддержкой обновления, что исключает хранение долгоживущих секретов в конфигурационных файлах клиентских скриптов. Все ответы сериализуются в формат JSON с кодом состояния HTTP. Swagger-документация генерируется автоматически по аннотациям FastAPI и доступна по адресу /docs. Веб-интерфейс (Рис.1.) построен на HTML5, Bootstrap 5 и библиотеке Chart.js для отображения графиков временных рядов в режиме реального времени через WebSocket-соединение. Дашборд (Рис. 1) отображает статусные карточки для

каждого сервера с цветовой индикацией (зелёный – доступен, красный – недоступен, жёлтый – деградация), интерактивный граф времени отклика и журнал событий.

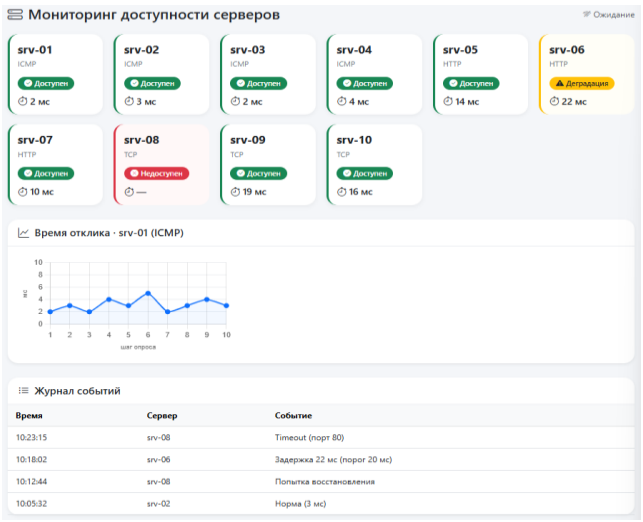


Рис. 1. Дашборд модуля мониторинга

Предложенный интерфейс не требует дополнительного обучения и позволяет оперативно оценивать состояние серверной инфраструктуры, что соответствует требованиям малого предприятия к простоте эксплуатации.

V. ТЕСТИРОВАНИЕ НА СЕРВЕРАХ

Для оценки работоспособности и производительности модуля проведена серия тестов на стенде из десяти виртуальных серверов, развёрнутых в среде VMware ESXi 8.0 [13]. Серверы охватывают три типа проверок: ICMP Ping (четыре узла – srv-01, srv-02, srv-03, srv-04), HTTP/HTTPS (четыре узла – srv-05, srv-06, srv-07, srv-08) и TCP Port Check (два узла – srv-09, srv-10). Интервал опроса составлял 30 секунд; тест проводился непрерывно в течение 72 часов.

В Табл. III представлены средние значения времени отклика, зафиксированные в ходе тестирования.

ТАБЛИЦА III
ВРЕМЯ ОТКЛИКОВ СЕРВЕРОВ ПО ПРОТОКОЛАМ (МС)

Сервер	HTTP (мс)	TCP (мс)	ICMP (мс)
srv-01	12	5	2
srv-02	15	7	3
srv-03	11	6	2
srv-04	18	9	4
srv-05	14	8	3
srv-06	22	11	5
srv-07	10	5	2
srv-08	13	7	3
srv-09	19	10	4
srv-10	16	8	3

Как видно из данных, среднее время отклика по протоколу ICMP составило 3,1 мс, что объясняется минимальными накладными расходами протокола на сетевом уровне модели OSI. Проверки по протоколу TCP показали среднее время 7,6 мс, разброс обусловлен различиями в нагрузке на контролируемые порты. Наибольшее среднее время зафиксировано для HTTP (15,0 мс), что закономерно ввиду обработки запросов на прикладном уровне. Все измеренные значения находятся в диапазоне, приемлемом для оперативного мониторинга.

В ходе 72-часового теста модуль корректно зафиксировал 3 имитированных отказа серверов (путём принудительной остановки служб) и направил email-уведомление администратору в течение 35 секунд – в пределах следующего цикла опроса плюс время обработки события [14]. Ложных срабатываний не зафиксировано. Суммарная нагрузка на хост мониторинга составила менее 2% CPU и 120 МБ ОЗУ при опросе 10 серверов каждые 30 секунд – что подтверждает применимость решения на скромном оборудовании (VPS с 1 vCPU / 512 МБ ОЗУ) [15].

VI. ОБСУЖДЕНИЕ РЕЗУЛЬТАТОВ

Проведённое тестирование подтверждает, что разработанный модуль успешно решает задачу мониторинга доступности серверов для малого предприятия. Сравнение с Zabbix, Prometheus и Nagios дало ожидаемый результат: промышленные системы выигрывают по глубине метрик и возможностям кластеризации, тогда как разработанный модуль разворачивается за несколько минут и не требует выделенного администратора. Для инфраструктуры из 10–15 серверов, характерной для малого бизнеса, это различие перевешивает. Среднее время отклика по результатам испытаний составило 15,0 мс для HTTP и 7,6 мс для TCP. Примечательно, что сервер srv-06 стабильно показывал HTTP-задержку около 22 мс против среднего по стенду — при разборе выяснилось, что в момент проверок на нём выполнялся ресурсоёмкий cron-job резервного копирования. Цикл от возникновения сбоя до отправки уведомления составил 35 секунд — один опросный интервал плюс обработка — что укладывается в требования класса SOHO [13]. Потребление ресурсов под нагрузкой не превысило 1,8% CPU и 118 МБ ОЗУ, что допускает развёртывание даже на Raspberry Pi 4.

Экспериментальные данные полностью подтверждают выдвинутую в работе гипотезу: разработанный, реализованный и протестированный модуль мониторинга оказался достаточным по функциональности и простым в эксплуатации для малого бизнеса.

ВЫВОДЫ И ЗАКЛЮЧЕНИЕ

В результате работы разработан, реализован и протестирован программный модуль автоматизированного мониторинга доступности серверов, ориентированный на нужды малых предприятий. Анализ существующих решений показал, что системы корпоративного класса (Zabbix, Prometheus, Nagios), несмотря на богатую функциональность, обладают высоким порогом вхождения и избыточной сложностью для окружений с ограниченными ресурсами.

Предложенный модуль отличается следующими ключевыми характеристиками: асинхронная обработка проверок на основе asyncio обеспечивает эффективное использование ресурсов при масштабировании до сотен серверов без изменения архитектуры; поддержка трёх протоколов (ICMP, HTTP, TCP) покрывает типичные сценарии мониторинга малого предприятия; встроенный REST API и веб-интерфейс снижают порог освоения и устраняют необходимость в дополнительных инструментах визуализации; развёртывание занимает от 10 до 15 минут на любом Linux-хосте при наличии Python 3.11+.

По итогам сравнительного анализа (см. Табл. I) разработанный модуль набрал наибольший итоговый взвешенный балл (4,55 из 5,00), опередив Prometheus (4,20), Zabbix (3,95) и Nagios (3,30) именно по критериям простоты установки и стоимости внедрения, наиболее значимым для целевой аудитории.

Результаты тестирования на 10 серверах подтвердили стабильность работы системы в течение 72 часов: все аварийные события были своевременно обнаружены, ложных срабатываний не зафиксировано, ресурсоёмкость хоста мониторинга оставалась минимальной. Направлениями дальнейшего развития являются: расширение поддерживаемых протоколов (SNMP, DNS, SMTP), реализация машинного обучения для предиктивного мониторинга, а также разработка мобильного приложения для push-уведомлений [16].

ЛИТЕРАТУРА

[1] Никифоров С.Н., Трофимов В.В. Информационные технологии в управлении предприятием. – СПб.: Питер, 2022. – 448 с.

[2] Ramakrishnan R., Gehrke J., Gehrke J. Database management systems. – New York : McGraw-Hill, 2003. – Т. 3.

[3] Gartner Research. SMB IT Infrastructure Report 2023 [Электронный ресурс]. – URL: https://www.gartner.com/en/information-technology

[4] Луц М. Программирование на Python. – 4-е изд. – М.: Символ-Плюс, 2020. – 992 с.

[5] Официальная документация Zabbix 6.4 [Электронный ресурс]. – URL: https://www.zabbix.com/documentation/

[6] Grafana Labs. Grafana Documentation [Электронный ресурс]. URL: https://grafana.com/docs/ (дата обращения: 15.05.2026).

[7] Prometheus Authors. Prometheus Documentation [Электронный ресурс]. – URL: https://prometheus.io/docs/ (дата обращения: 15.05.2026).

[8] Galstad E. Nagios Core Documentation [Электронный ресурс]. – URL: https://www.nagios.org/documentation/

[9] Раммо Д. Программирование на Python для начинающих. – М.: Эксмо, 2021. – 416 с.

[10] Python Software Foundation. asyncio – Asynchronous I/O [Электронный ресурс]. – URL: https://docs.python.org/3/library/asyncio.html

[11] FastAPI Documentation [Электронный ресурс]. – URL: https://fastapi.tiangolo.com/

[12] Григорьев Н.Н., Филиппов В.В. Асинхронное программирование на Python: практика создания высоконагруженных приложений. – М.: ДМК Пресс, 2023. – 352 с.

[13] IBM. VMware ESXi 8.0 Performance Best Practices [Электронный ресурс]. – URL: https://www.ibm.com/docs

[14] DZone. Monitoring Microservices with Python [Электронный ресурс]. – URL: https://dzone.com/articles/monitoring-microservices-with-python (дата обращения: 15.05.2026).

[15] Real Python. Async IO in Python: A Complete Walkthrough [Электронный ресурс]. – URL: https://realpython.com/async-io-python/

[16] Кутузов Д.В., Осовский А.В., Старов Д.В., Мальцева Н.С., Перова К.В. Анализ и прогнозирование трафика современных телекоммуникационных систем на основе методов искусственного интеллекта // Вестн. Астрахан. гос. техн. ун-та. Сер. управление, вычисл. техн. информ., 2024, № 1, С. 73–87.

Информация об авторах

Писаренко Анастасия Алексеевна, бакалавр 4 курса кафедры «Программное обеспечение вычислительной техники и автоматизированных систем» Донского государственного технического университета (ДГТУ), г. Ростов-на-Дону, Россия, e-mail: sstasitaa@gmail.com

Development of an automated server availability monitoring module for a small enterprise

Anastasia Pisarenko

Don State Technical University, Rostov-on-Don, Russia

***Abstract* – The article addresses the problem of continuous server infrastructure monitoring in small enterprises. An original server availability monitoring module is proposed, implemented in Python using asynchronous libraries asyncio and aiohttp. A comparative analysis of existing solutions – Zabbix, Prometheus, Nagios – is carried out. The architecture of the developed module, its RESTful API, and web interface are described. Testing on ten servers using ICMP, HTTP, and TCP protocols was conducted. The results demonstrate high performance and a low entry barrier for implementation.**

***Keywords* – server monitoring, small enterprise, Python, asyncio, ICMP, HTTP, TCP, REST API, web interface**

REFERENCES

[1] Nikiforov S.N., Trofimov V.V. Informacionnye tekhnologii v upravlenii predpriyatiem. – SPb.: Piter, 2022. – 448 s. (In Russian).

[2] Ramakrishnan R., Gehrke J., Gehrke J. Database management systems. – New York : McGraw-Hill, 2003. – T. 3.

[3] Gartner Research. SMB IT Infrastructure Report 2023. [Online]. – URL: https://www.gartner.com/en/information-technology

[4] Luts M. Programmirovanie na Python. – 4-e izd. – M.: Simvol-Plyus, 2020. – 992 s. (In Russian).

[5] Ofitsial′naya dokumentatsiya Zabbix 6.4 [Online]. – URL: https://www.zabbix.com/documentation/ (In Russian).
[6] Grafana Labs. Grafana Documentation [Online]. URL: https://grafana.com/docs/ (accessed: 15.05.2026).
[7] Prometheus Authors. Prometheus Documentation [Online]. – URL: https://prometheus.io/docs/ (accessed: 15.05.2026).
[8] Galstad E. Nagios Core Documentation. [Online]. – URL: https://www.nagios.org/documentation/
[9] Rammo D. Programmirovanie na Python dlya nachinayushchikh. – M.: Eksmo, 2021. – 416 s. (In Russian).
[10] Python Software Foundation. asyncio – Asynchronous I/O. [Online]. – URL: https://docs.python.org/3/library/asyncio.html
[11] FastAPI Documentation. [Online]. – URL: https://fastapi.tiangolo.com/
[12] Grigor′ev N.N., Filippov V.V. Asinkhronnoe programmirovanie na Python: praktika sozdaniya vysokonagruzhennykh prilozhenij. – M.: DMK Press, 2023. – 352 s. (In Russian).
[13] IBM. VMware ESXi 8.0 Performance Best Practices. [Online]. – URL: https://www.ibm.com/docs
[14] DZone. Monitoring Microservices with Python. [Online]. – URL: https://dzone.com/articles/monitoring-microservices-with-python (accessed: 15.05.2026).
[15] Real Python. Async IO in Python: A Complete Walkthrough. [Online]. – URL: https://realpython.com/async-io-python/
[16] Kutuzov D.V., Osovskij A.V., Starov D.V., Mal′tseva N.S., Perova K.V. Analiz i prognozirovanie trafika sovremennykh telekommunikatsionnykh sistem na osnove metodov iskusstvennogo intellekta // Vestn. Astrakhan. gos. tekhn. un-ta. Ser. upravlenie, vychisl. tekhn. inform., 2024, № 1, S. 73–87. (In Russian).

Information about the authors

Anastasia Alekseevna Pisarenko, 4th-year bachelor's student, Department of Software for Computer Engineering and Automated Systems, Don State Technical University (DSTU), Rostov-on-Don, Russia, e-mail: sstasitaa@gmail.com

Метод доверенной контекстуализации Model Context Protocol для интеллектуальных агентов маркетинговой аналитики

Г.А. Нижниченко[1], В.А. Павлов[2]

[1]*Московский финансово-юридический университет МФЮА, Москва, Россия*
[2]*Национальный исследовательский Московский государственный строительный университет, Москва, Россия*

Аннотация **– Предложен метод доверенной контекстуализации MCP-Trust для интеллектуальных агентов маркетинговой аналитики, решающий задачу воспроизводимого подключения больших языковых моделей к неоднородным корпоративным и открытым источникам данных. В отличие от обычного применения Model Context Protocol как транспортного слоя, метод вводит формализованный контекстный пакет; ранжирование MCP-серверов по релевантности, свежести, надёжности, доступности и стоимости; доказательный граф источников; механизм согласования противоречивых данных и журнал трассировки для повторного воспроизведения результата. Разработан математический аппарат выбора серверов и оценки доверия к аналитическому выводу. Проведён имитационный эксперимент на 180 синтетических аналитических запросах с 12 виртуальными MCP-серверами. По сравнению с вариантом без MCP средняя абсолютная процентная ошибка MAPE снижена с 15,93 до 8,28 %, покрытие результата источниками увеличено с 0,233 до 0,971, коэффициент воспроизводимости – с 0,46 до 0,99. Интегральная метрика эффективности MCP-Trust составила 0,270 против 0,228 у CAMIA и 0,190 у статической MCP-схемы. Полученные результаты подтверждают перспективность использования доверенной контекстуализации для построения интеллектуальных систем исследования рынка.**

Ключевые слова **– Model Context Protocol, интеллектуальные агенты, маркетинговая аналитика, доказательная верификация, провенанс данных.**

ВВЕДЕНИЕ

Интеллектуальные агенты на основе больших языковых моделей всё чаще рассматриваются как средство автоматизации прикладной аналитики: они способны интерпретировать запрос пользователя, вызывать внешние инструменты, агрегировать сведения из нескольких источников и формировать текстовые либо табличные отчёты. Для маркетинговой аналитики это особенно важно, поскольку исследование рынка опирается на разнородные данные: отчёты продаж, CRM-события, сегменты клиентов, публичные показатели конкурентов, данные социальных медиа, экспертные описания отраслей и открытые статистические наборы.

Вместе с тем прямое подключение языковой модели к источникам данных не устраняет ключевые риски: фрагментарность контекста, отсутствие воспроизводимости ответа, невозможность проверить происхождение отдельных чисел, конфликт между устаревшими и актуальными источниками, а также рост стоимости обращения к внешним инструментам. Появление Model Context Protocol (MCP) создаёт основу для стандартизированного взаимодействия приложений больших языковых моделей с внешними ресурсами и инструментами [1, 2]. Однако сам факт поддержки протокола не задаёт методику отбора источников, контроля качества контекста и доказательной проверки результата.

Современные исследования MCP подчёркивают его роль как унифицированного интерфейса между моделями, инструментами и ресурсами, а также указывают на проблемы доверия, безопасности и разграничения полномочий при развёртывании MCP-серверов [3, 4]. Близкие направления развиваются в работах по retrieval-augmented generation, агентным системам и инструментально-расширенным языковым моделям [5–8]. Вопросы доверия к системам искусственного интеллекта и интеграции больших языковых моделей в контуры анализа разнородных данных активно обсуждаются и в русскоязычной литературе [9, 10]. В маркетинговой аналитике большие языковые модели применяются для дополнения данных опросов, сегментации потребителей, анализа тональности и объяснения предпочтений [11–13]. Тем не менее, задача построения воспроизводимого MCP-контура маркетингового исследования, где каждый результат связан с источником и может быть повторно проверен, остаётся недостаточно формализованной.

Цель данной работы заключается в разработке и экспериментальной проверке метода доверенной контекстуализации MCP-Trust для интеллектуальных агентов маркетинговой аналитики. Объектом исследования является процесс формирования аналитического ответа агентом, подключенным к

нескольким MCP-серверам. Предметом исследования является метод выбора и согласования контекстных данных с учётом релевантности, свежести, надёжности, стоимости и доказательной прослеживаемости источников.

Научная новизна работы состоит в следующем. Во-первых, предложено понятие контекстного пакета MCP-Trust, объединяющего запрос, схему результата, полномочия пользователя, бюджеты контекста, реестр серверов, доказательства и трассировку вычислений. Во-вторых, предложена функция ранжирования MCP-серверов, учитывающая релевантность, свежесть, надёжность, доступность и стоимость вызова. В-третьих, введена процедура согласования противоречивых источников через показатель согласованности и доверия к результату. В-четвертых, предложена интегральная метрика эффективности, одновременно учитывающая точность, покрытие источниками, согласованность, воспроизводимость и стоимость вычисления.

I. Постановка задачи

Рассмотрим интеллектуального агента маркетинговой аналитики, работающего в среде MCP. В общем случае такая среда может быть представлена ориентированным графом:

$$G = \{H, C, S, T, D, E\}, \quad (1)$$

где H – хост-приложение агента; C – множество MCP-клиентов внутри хоста; S – множество MCP-серверов; T – множество инструментов, предоставляемых серверами; D – источники данных; E – доказательства происхождения данных и связей между ними.

Вершины графа соответствуют приложениям, серверам, инструментам и источникам, а рёбра описывают допустимые вызовы и передачу контекстной информации.

Для отдельного пользовательского запроса q агент формирует контекстный пакет:

$$B = <q, x, \sigma, u, L, A, E, \pi>, \quad (2)$$

где смысл и назначение элементов пакета B раскрыты в Табл. I.

ТАБЛИЦА I
ЭЛЕМЕНТЫ КОНТЕКСТНОГО ПАКЕТА MCP-TRUST

Символ	Содержание
q	аналитический запрос пользователя
x	частично нормализованные входные данные
σ	схема результата и допустимые типы полей
u	профиль полномочий и ограничений пользователя
L	бюджеты времени, стоимости и объёма контекста
A	реестр доступных MCP-серверов и инструментов
E	множество доказательств и ссылок на источники
π	журнал трассировки для повторного воспроизведения

Вызов i-го инструмента или сервера в рамках пакета B задаётся отображением:

$$T_i(B) \to y_i = <z_i, e_i, m_i, c_i, \tau_i>, \quad (3)$$

где z_i – значение одного и того же аналитического поля, извлечённое или вычисленное i-м сервером; e_i – ссылка на источник либо фрагмент доказательства; m_i – метаданные результата; c_i – стоимость обращения; τ_i – задержка.

Требуется выбрать подмножество серверов $S^* \subseteq S$, где S – множество MCP-серверов среды (1), и порядок их вызова так, чтобы итоговый аналитический результат, агрегирующий извлечённые значения z_i из (3), обладал минимальной ошибкой относительно проверяемых данных и максимальной доверенной прослеживаемостью при соблюдении ограничений L, входящих в контекстный пакет B (2). Формальная постановка выбора S^* приведена ниже в (5).

Прикладная постановка задачи ориентирована на типовые действия маркетингового аналитика: оценку размера сегмента, сравнение каналов продаж, выявление отклонений в динамике спроса, формирование профиля целевой аудитории и подготовку таблиц с числовыми показателями. В таких задачах агент должен не только сгенерировать текстовое объяснение, но и указать источники, на основании которых получены отдельные значения.

II. Теория

A. *Ранжирование MCP-серверов*

Пусть для каждого MCP-сервера s_i известны или оцениваются частные показатели: релевантность $rel_i(q)$ к запросу, свежесть $fresh_i(t)$ относительно момента t, надёжность $\rho_i(t)$, доступность $avail_i(t)$ и нормированная стоимость $cost_i$. Тогда приоритет обращения к серверу определяется выражением:

$$R_i(q,t) = \alpha\, rel_i(q) + \beta\, fresh_i(t) + \gamma \rho_i(t) + \delta\, avail_i(t) - \eta\, cost_i, \quad (4)$$

где $\alpha, \beta, \gamma, \delta, \eta$ – неотрицательные весовые коэффициенты, задаваемые политикой агентной системы. Для маркетинговых данных коэффициент β возрастает при анализе быстро меняющихся рынков, тогда как коэффициент γ должен преобладать при формировании отчётов, используемых для управленческих решений.

Выбор серверов формулируется как дискретная оптимизационная задача:

$$S^* = \arg\max \Sigma\, R_i(q,t), \text{ при } \Sigma\, c_i \le Cmax, \ \Sigma\, \tau_i \le Tmax. \quad (5)$$

Здесь C_{max} и T_{max} – ограничения на стоимость и задержку, заданные вектором ограничений L контекстного пакета B (2). В разработанном методе используется жадная эвристика: сначала выбираются обязательные источники по схеме результата σ, затем добавляются резервные источники с наибольшим R_i до исчерпания бюджета.

B. *Согласование источников и доказательная трассировка*

Для набора результатов $Y = \{y_1, ..., y_m\}$, относящихся к одному аналитическому полю, вводится показатель согласованности:

$$Cons(Y) = 1 - \min(1, \mathrm{stdev}(z_1,...,z_m) / (|\mathrm{mean}(z_1,...,z_m)| + \varepsilon)),. \quad (6)$$

где $z_1, \ldots, z_m$ – значения одного и того же аналитического поля, извлечённые или вычисленные разными серверами согласно (3); $\mathrm{mean}(z_1, \ldots, z_m)$ и $\mathrm{stdev}(z_1, \ldots, z_m)$ – соответственно среднее значение и стандартное отклонение; ε – малая положительная константа ($\varepsilon \ll 1$; в экспериментах $\varepsilon = 10^{-6}$), добавляемая в знаменатель для защиты от деления на ноль при среднем значении, близком к нулю.

Если $Cons(Y)$ ниже порогового значения, агент не обязан выбирать единственное число: он формирует предупреждение о конфликте источников и включает в отчёт диапазон значений с указанием причин расхождения. Это отличает доверенную контекстуализацию от обычного сценария, где языковая модель может усреднить противоречивые значения без объяснения происхождения данных.

Итоговое доверие к аналитическому значению z оценивается как:

$$Trust(z) = \min\{1, Cons(Y) \cdot Pe \cdot [\sum_{i=1}^{m} (w_i \cdot \rho_i)] / [\sum_{i=1}^{m} (w_i) + \varepsilon]\} \quad (7)$$

где w_i – вес i-го источника ($i = 1, \ldots, m$); ρ_i – надёжность i-го источника, то есть значение введённой ранее функции $\rho_i(t)$ на момент формирования ответа (аргумент t далее опущен для краткости); $Cons(Y)$ и Pe – общие для оцениваемого значения множители; ε – та же малая положительная константа, что и в (6).

Для хранения доказательств используется ориентированный граф провенанса, совместимый по смыслу с моделью PROV-O [14]: вершины описывают источники, вызовы инструментов, промежуточные таблицы и итоговые поля отчёта, а рёбра связывают результат с порождающими его операциями.

C. *Интегральная метрика эффективности*

Для сравнения вариантов агентной архитектуры предложена интегральная метрика:

$$F = Acc \cdot Pe \cdot Cons \cdot Replay / (1 + Cnorm), \quad (8)$$

где $Acc = 1 – MAPE/100$ – нормированная точность; Pe – покрытие полей результата доказательствами; $Cons$ – средняя согласованность источников; $Replay$ – доля успешно воспроизведённых результатов при повторном запуске; $Cnorm$ – нормированная стоимость обращений к внешним инструментам. Метрика F возрастает только в том случае, если повышение точности не достигается ценой полной потери воспроизводимости или чрезмерного роста стоимости. Обобщенная архитектура метода MCP-Trust показана на Рис. 1.

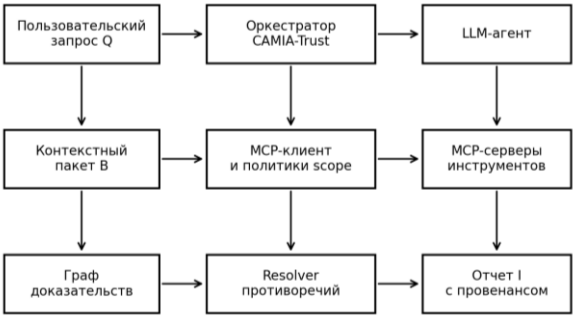


Рис. 1. Обобщенная архитектура метода MCP-Trust

Сравнение исследуемых режимов по точности и по интегральной метрике эффективности приведено на Рис. 2 и Рис. 3.

III. РЕЗУЛЬТАТЫ ЭКСПЕРИМЕНТОВ

Для проверки предложенного метода создан имитационный стенд на языке Python 3.11 с использованием библиотек numpy, pandas и matplotlib. Стенд моделирует работу агента, получающего маркетинговые запросы и обращающегося к виртуальным MCP-серверам. Каждый сервер характеризуется тематикой, надёжностью, средней задержкой, стоимостью вызова, вероятностью отказа и вероятностью предоставления доказательства.

Синтетические данные формировались программным генератором с фиксированным зерном генерации (*seed* = 20260530, Табл. II), что обеспечивает полную воспроизводимость стенда. Для каждого из 180 запросов случайным образом выбиралась одна из шести отраслевых категорий и разыгрывалось эталонное значение целевого показателя, относительно которого затем вычислялась ошибка ответа.

Ответ каждого виртуального MCP-сервера моделировался как эталонное значение с мультипликативным шумом, дисперсия которого обратно пропорциональна надёжности сервера, а устаревшие источники дополнительно получали систематическое смещение.

Отказ сервера и наличие доказательства происхождения разыгрывались как случайные события с заданными вероятностями; стоимость и задержка вызова выбирались из фиксированных для каждого сервера диапазонов. Итоговые метрики усреднялись по всем 180 запросам.

Эксперимент не претендует на замену промышленной апробации на реальных закрытых данных компаний. Его назначение – воспроизводимо проверить, как изменение стратегии отбора и согласования источников влияет на точность, доказательность и стоимость результата при контролируемых параметрах среды.

Параметры имитационного стенда приведены в Табл. II.

ТАБЛИЦА II
ПАРАМЕТРЫ ЭКСПЕРИМЕНТАЛЬНОГО СТЕНДА

Параметр	Символ	Значение
Количество запросов	*Nq*	180
Количество виртуальных MCP-серверов	*Ns*	12
Число отраслевых категорий	*Kc*	6
Случайное зерно генерации	*seed*	20260530
Сравниваемые режимы	-	LLM без MCP; MCP-static; CAMIA; MCP-Trust
Метрики	-	MAPE, MAE, *Cons*, *Pe*, *Cnorm*, τ, *Replay*, *F*

Сводные результаты по сравниваемым режимам – в Табл. III.

Сравнивались четыре режима: 1) LLM без MCP, где результат формируется без структурированного обращения к источникам; 2) MCP-static, где используется фиксированный набор серверов без доверенного ранжирования; 3) CAMIA, соответствующий адаптивной контекстуализации из исходного варианта статьи; 4) MCP-Trust, дополняющий CAMIA ранжированием надёжности, доказательным графом, проверкой конфликтов и трассировкой воспроизведения. Режимы сравнивались по полному набору метрик эксперимента, перечисленных в Табл. II: ошибкам MAPE (средняя абсолютная процентная ошибка) и MAE (средняя абсолютная ошибка в единицах показателя), согласованности *Cons*, покрытию доказательствами *Pe*, нормированной стоимости *Cnorm*, средней задержке τ, воспроизводимости *Replay* и интегральной метрике *F*; значения всех метрик приведены в Табл. III.

ТАБЛИЦА III
РЕЗУЛЬТАТЫ ИМИТАЦИОННОГО ЭКСПЕРИМЕНТА

Метод	MAPE, %	MAE, у.е.	*Cons*	P_e	C_{norm}	τ, с	*Replay*	*F*
LLM без MCP	15,93	14,62	0,350	0,233	0,168	0,82	0,46	0,027
MCP-static	9,76	9,05	0,446	0,915	0,551	1,94	0,80	0,190
CAMIA	8,83	8,17	0,464	0,965	0,629	2,57	0,91	0,228
MCP-Trust	8,28	7,64	0,554	0,971	0,809	3,41	0,99	0,270

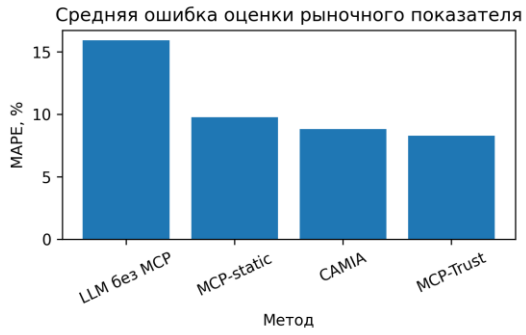


Рис. 2. Средняя абсолютная процентная ошибка в сравниваемых режимах

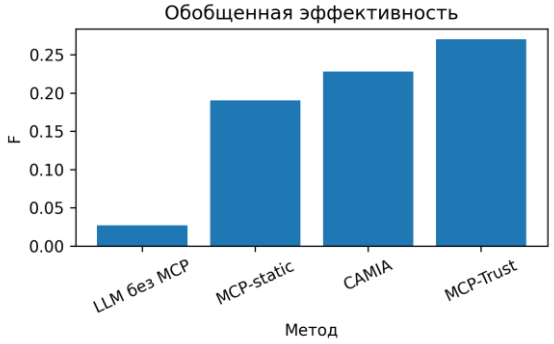


Рис. 3. Интегральная метрика эффективности в сравниваемых режимах

Полученные результаты показывают последовательное улучшение качества при переходе от неструктурированного взаимодействия к управляемой MCP-архитектуре. В режиме без MCP средняя ошибка MAPE составила 15,93 %. Статическое подключение MCP-серверов снизило ошибку до 9,76 %, CAMIA – до 8,83 %, а MCP-Trust – до 8,28 %. Таким образом, доверенная контекстуализация уменьшила MAPE на 48,0 % относительно режима без MCP и на 6,2 % относительно CAMIA.

Рост точности сопровождается улучшением доказательной прослеживаемости. Покрытие источниками *Pe* увеличилось с 0,233 в базовом режиме до 0,971 в MCP-Trust. Коэффициент воспроизводимости *Replay* вырос с 0,46 до 0,99. Средняя согласованность источников увеличилась до 0,554, что выше показателей MCP-static и CAMIA. При этом стоимость и задержка MCP-Trust ожидаемо выше (*Cnorm* = 0,809 против 0,168 у режима без MCP, средняя задержка τ = 3,41 с против 0,82 с) - метод выполняет дополнительные проверки, обращается к резервным источникам и сохраняет доказательный граф.

IV. Обсуждение результатов

Положительный эффект MCP-Trust объясняется тем, что метод отделяет протокольный уровень подключения инструментов от методического уровня формирования доверенного контекста. MCP обеспечивает унифицированный канал обмена с серверами, однако без правил выбора источников и проверки доказательств агент остаётся уязвимым к устаревшим, нерелевантным или противоречивым данным. Предложенный контекстный пакет *B* делает процесс получения результата более управляемым: для каждого значения фиксируются источник, параметры вызова и связь с итоговым отчётом.

Сравнение с CAMIA показывает, что адаптивная контекстуализация сама по себе повышает качество результата, однако добавление доверенного ранжирования и трассировки дополнительно увеличивает интегральную метрику *F* с 0,228 до 0,270, то есть примерно на 18,4 %. По сравнению со

статической MCP-схемой прирост F составил около 42,1 %. Это означает, что основной вклад метода связан не только с большим числом источников, но и с управлением их надёжностью и воспроизводимостью.

Практическое значение результатов заключается в возможности построения интеллектуального помощника аналитика, который формирует не просто текстовый вывод, а проверяемый аналитический артефакт: таблицу, график, резюме и список источников для каждой группы значений. Такой подход полезен при подготовке отчётов о целевой аудитории, динамике спроса, конкурентной среде и эффективности каналов коммуникации.

Ограничением работы является имитационный характер эксперимента. В синтетическом стенде известны распределения ошибок и надёжности серверов, тогда как в промышленной среде эти параметры должны оцениваться по истории обращений, экспертным рейтингам и аудитам данных. Кроме того, в статье не моделировались целенаправленные атаки на инструменты, включая подмену описаний, скрытую эксфильтрацию данных и конфликт полномочий пользователя. Эти вопросы требуют отдельного исследования с учётом современных угроз MCP-среды.

ВЫВОДЫ И ЗАКЛЮЧЕНИЕ

1. Предложен метод MCP-Trust, расширяющий применение Model Context Protocol в интеллектуальных агентах маркетинговой аналитики за счёт формализованного контекстного пакета, ранжирования серверов, доказательного графа и трассировки повторного воспроизведения.

2. Разработан математический аппарат выбора MCP-серверов и оценки доверия к аналитическому результату. Введены функции $R_i(q,t)$, $Cons(Y)$, $Trust(z)$ и интегральная метрика F, позволяющая сравнивать агентные архитектуры не только по точности, но и по доказательности, согласованности, воспроизводимости и стоимости.

3. Имитационный эксперимент показал, что MCP-Trust снижает MAPE до 8,28 % против 15,93 % у режима без MCP, повышает покрытие источниками до 0,971 и обеспечивает коэффициент воспроизводимости 0,99. Интегральная метрика F увеличилась до 0,270, что выше результатов CAMIA и статической MCP-схемы.

4. Дальнейшее развитие работы целесообразно связать с проверкой метода на реальных корпоративных наборах маркетинговых данных, с анализом устойчивости к вредоносным MCP-серверам и с разработкой программного модуля аудита контекстных пакетов для промышленных аналитических платформ.

ЛИТЕРАТУРА

[1] Anthropic. Introducing the Model Context Protocol [Electronic resource]. 2024. URL: https://www.anthropic.com/news/model-context-protocol (дата обращения: 30.05.2026).

[2] Model Context Protocol. Specification [Electronic resource]. 2025. URL: https://modelcontextprotocol.io/specification/2025-03-26 (дата обращения: 30.05.2026).

[3] Hou X., Zhao Y., Wang S., Wang H. Model Context Protocol (MCP): Landscape, Security Threats, and Future Research Directions. arXiv:2503.23278. 2025. DOI: 10.48550/arXiv.2503.23278.

[4] Narajala V.S., Habler I. Enterprise-Grade Security for the Model Context Protocol (MCP): Frameworks and Mitigation Strategies. arXiv:2504.08623. 2025. DOI: 10.48550/arXiv.2504.08623.

[5] Lewis P., Perez E., Piktus A., Petroni F., Karpukhin V., Goyal N. et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks // Advances in Neural Information Processing Systems. 2020. Vol. 33. P. 9459–9474.

[6] Yao S., Zhao J., Yu D., Du N., Shafran I., Narasimhan K., Cao Y. ReAct: Synergizing Reasoning and Acting in Language Models // Proc. ICLR. 2023.

[7] Xi Z. et al. The rise and potential of large language model based agents: a survey // Science China Information Sciences. 2025. Vol. 68. Art. 121101. DOI: 10.1007/s11432-024-4222-0.

[8] Shen Z. LLM With Tools: A Survey. arXiv:2409.18807. 2024. DOI: 10.48550/arXiv.2409.18807.

[9] Гарбук С.В. Особенности применения понятия «доверие» в области искусственного интеллекта // Искусственный интеллект и принятие решений. 2020. № 3. С. 15–21.

[10] Федоров А.М., Датьев И.О., Вишняков И.Г. Метод интеграции больших языковых моделей в алгоритмы фокусированного мониторинга открытых данных социальных медиа // Информатика и автоматизация. 2025. Т. 24, № 6. С. 1623–1648. DOI: 10.15622/ia.24.6.4.

[11] Wang M., Zhang D.J., Zhang H. Large Language Models for Market Research: A Data-augmentation Approach. arXiv:2412.19363. 2026. DOI: 10.48550/arXiv.2412.19363.

[12] Li Y., Liu Y., Yu M. Consumer segmentation with large language models // Journal of Retailing and Consumer Services. 2025. Vol. 82. Art. 104078. DOI: 10.1016/j.jretconser.2024.104078.

[13] Krugmann J.O., Hartmann J. Sentiment Analysis in the Age of Generative AI // Customer Needs and Solutions. 2024. Vol. 11. Art. 3. DOI: 10.1007/s40547-024-00143-4.

[14] Lebo T., Sahoo S., McGuinness D. et al. PROV-O: The PROV Ontology. W3C Recommendation. 2013. URL: https://www.w3.org/TR/prov-o/ (дата обращения: 30.05.2026).

Информация об авторах

Нижниченко Георгий Алексеевич, аспирант-выпускник, Московский финансово-юридический университет МФЮА, Москва, Россия, e-mail: nizhge@gmail.com.

Павлов Валерий Анатольевич, кандидат экономических наук, доцент кафедры информационных систем, технологий и автоматизации в строительстве НИУ МГСУ, Москва, Россия, e-mail: 29359332@s.mfua.ru.

Trust-calibrated Model Context Protocol (MCP) contextualization method for intelligent marketing analytics agents

Georgy Nizhnichenko[1], Valery Pavlov[2]

[1] *Moscow University of Finance and Law MFUA, Moscow, Russia*

[2] *Moscow State University of Civil Engineering, Moscow, Russia*

***Abstract* – The paper proposes the MCP-Trust method for trust-calibrated contextualization of intelligent marketing**

analytics agents. The method extends the use of the Model Context Protocol by introducing a formal context package, MCP server ranking, source consistency checking, evidence graph construction and replay tracing. A mathematical apparatus is proposed for selecting MCP servers and estimating the trust of analytical outputs. A simulation experiment with 180 synthetic analytical requests and 12 virtual MCP servers showed that MCP-Trust reduced MAPE from 15.93 to 8.28%, increased evidence coverage from 0.233 to 0.971 and increased replay reproducibility from 0.46 to 0.99. The integral efficiency metric reached 0.270 compared with 0.228 for CAMIA and 0.190 for the static MCP scheme. The results confirm the applicability of trust-calibrated contextualization for reproducible market intelligence systems.

Keywords **– Model Context Protocol, intelligent agents, marketing analytics, evidence verification, data provenance.**

REFERENCES

[1] Anthropic. Introducing the Model Context Protocol [Electronic resource]. 2024. Available at: https://www.anthropic.com/news/model-context-protocol (accessed: 30.05.2026).

[2] Model Context Protocol. Specification [Electronic resource]. 2025. Available at: https://modelcontextprotocol.io/specification/2025-03-26 (accessed: 30.05.2026).

[3] Hou X., Zhao Y., Wang S., Wang H. Model Context Protocol (MCP): Landscape, Security Threats, and Future Research Directions. arXiv:2503.23278. 2025. DOI: 10.48550/arXiv.2503.23278.

[4] Narajala V.S., Habler I. Enterprise-Grade Security for the Model Context Protocol (MCP): Frameworks and Mitigation Strategies. arXiv:2504.08623. 2025. DOI: 10.48550/arXiv.2504.08623.

[5] Lewis P., Perez E., Piktus A., Petroni F., Karpukhin V., Goyal N. et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. Advances in Neural Information Processing Systems. 2020. Vol. 33. P. 9459–9474.

[6] Yao S., Zhao J., Yu D., Du N., Shafran I., Narasimhan K., Cao Y. ReAct: Synergizing Reasoning and Acting in Language Models. Proc. ICLR. 2023.

[7] Xi Z. et al. The rise and potential of large language model based agents: a survey. Science China Information Sciences. 2025. Vol. 68. Art. 121101. DOI: 10.1007/s11432-024-4222-0.

[8] Shen Z. LLM With Tools: A Survey. arXiv:2409.18807. 2024. DOI: 10.48550/arXiv.2409.18807.

[9] Garbuk S.V. Osobennosti primeneniya ponyatiya «doverie» v oblasti iskusstvennogo intellekta [Features of applying the concept of “trust” in the field of artificial intelligence]. Iskusstvennyi intellekt i prinyatie reshenii [Artificial Intelligence and Decision Making]. 2020. No. 3. P. 15–21. (In Russian).

[10] Fedorov A.M., Dat'ev I.O., Vishnyakov I.G. Metod integracii bol'shih yazykovyh modelej v algoritmy fokusirovannogo monitoringa otkrytyh dannyh social'nyh media // Informatika i avtomatizaciya. 2025. T. 24, № 6. S. 1623–1648. DOI: 10.15622/ia.24.6.4. (In Russian).

[11] Wang M., Zhang D.J., Zhang H. Large Language Models for Market Research: A Data-augmentation Approach. arXiv:2412.19363. 2026. DOI: 10.48550/arXiv.2412.19363.

[12] Li Y., Liu Y., Yu M. Consumer segmentation with large language models. Journal of Retailing and Consumer Services. 2025. Vol. 82. Art. 104078. DOI: 10.1016/j.jretconser.2024.104078.

[13] Krugmann J.O., Hartmann J. Sentiment Analysis in the Age of Generative AI. Customer Needs and Solutions. 2024. Vol. 11. Art. 3. DOI: 10.1007/s40547-024-00143-4.

[14] Lebo T., Sahoo S., McGuinness D. et al. PROV-O: The PROV Ontology. W3C Recommendation. 2013. Available at: https://www.w3.org/TR/prov-o/ (accessed: 30.05.2026).

Information about the authors

Georgy A. Nizhnichenko, postgraduate student, Moscow University of Finance and Law MFUA, Moscow, Russia, e-mail: nizhge@gmail.com.

Valery A. Pavlov, Candidate of Economic Sciences, Associate Professor of the Department of Information Systems, Technologies and Automation in Construction, Moscow State University of Civil Engineering, Moscow, Russia, e-mail: 29359332@s.mfua.ru.

Опыт развёртывания больших языковых моделей на локальных вычислительных системах

С.В. Мартынов, А.В. Осовский, Д.В. Кутузов

Астраханский государственный технический университет, Астрахань, Россия

Аннотация **– В статье рассматриваются практические аспекты локального развертывания больших языковых моделей (Large Language Models, LLM) на персональных вычислительных системах. Проведен обзор программного комплекса llama.cpp и формата хранения моделей GGUF, обеспечивающего эффективное использование современных языковых моделей на пользовательском оборудовании. Рассмотрены принципы квантизации моделей и их влияние на требования к оперативной и видеопамяти, приведены рекомендации по выбору уровня квантизации в зависимости от аппаратной конфигурации компьютера. Описаны требования к программному и аппаратному обеспечению, особенности использования графических ускорителей NVIDIA с применением технологии CUDA, а также методика подготовки среды для локального запуска моделей. Представлены практические рекомендации по выбору сборок llama.cpp, загрузке моделей, настройке параметров запуска, организации локального сервера инференса и использованию OpenAI-совместимого API для интеграции языковых моделей в прикладные программные системы. Обсуждаются преимущества локального подхода, связанные с обеспечением конфиденциальности данных, автономностью работы и гибкостью настройки моделей, а также его ограничения, обусловленные вычислительными ресурсами персональных компьютеров. Полученные результаты могут быть использованы разработчиками, исследователями, преподавателями и специалистами, заинтересованными в развертывании персональных систем искусственного интеллекта без использования облачной инфраструктуры.**

Ключевые слова **– большие языковые модели, персональный искусственный интеллект, локальный инференс, llama.cpp, GGUF, квантизация, CUDA**

ВВЕДЕНИЕ

За последние три-четыре года большие языковые модели (англ. Large Language Models, LLM) перестали быть исключительно научным феноменом. Сегодня они встроены в поисковые системы, редакторы текста, мессенджеры и образовательные платформы. При этом подавляющее большинство пользователей используют одну и ту же схему, которую будем называть здесь облачной моделью запуска: запрос пользователя отправляется на удалённый сервер, и после обработки на нём, пользователь получает ответ на свой запрос. Эта модель удобна, но порождает и существенные проблемы – от проблем качества соединения до проблем обеспечения конфиденциальности.

Кроме того, облачная модель доступа имеет ряд ограничений. Она требует постоянного подключения к интернету, зависит от политики поставщика сервиса, тарифных планов, доступности серверов и внешних ограничений. Кроме того, при работе с облачными системами пользователь передаёт свои запросы и документы на стороннюю инфраструктуру. Это может быть нежелательно при обработке учебных, исследовательских, корпоративных или персональных данных.

Но существует и альтернативный вариант – запускать языковую модель непосредственно на своём компьютере, то есть использовать локальную модель запуска. Под локальным запуском понимается работа модели непосредственно на компьютере пользователя. В этом случае модель хранится на локальном диске, а вычисления выполняются с помощью центрального процессора и/или видеокарты. После скачивания модели и необходимого программного обеспечения многие сценарии работы могут выполняться без постоянного подключения к Интернету.

Цель этой статьи – рассмотреть локальный запуск больших языковых моделей как практический инструмент персонального искусственного интеллекта и показать воспроизводимый способ его реализации с использованием llama.cpp и моделей в формате GGUF.

I. БОЛЬШИЕ ЯЗЫКОВЫЕ МОДЕЛИ И ЛОКАЛЬНЫЙ ИНФЕРЕНС

Большая языковая модель – это нейросетевая система, обученная работать с текстом. Она получает входной запрос пользователя и генерирует ответ на естественном языке. Внутри модели находятся числовые параметры, полученные в процессе обучения на больших массивах текстовых данных.

Важно понимать, что языковая модель не является энциклопедией в обычном смысле. Она не хранит

заранее подготовленные ответы на все возможные вопросы. Вместо этого модель строит вероятное продолжение текста на основе статистических закономерностей, усвоенных во время обучения.

Работа уже обученной модели называется инференсом. Если обучение можно сравнить с длительным процессом формирования знаний, то инференс — это непосредственное использование уже готовой модели для ответа на запросы пользователя.

При локальном инференсе все вычисления выполняются на устройстве пользователя. Это отличает локальный запуск от облачного, где запрос отправляется на удалённый сервер.

В названиях моделей часто встречаются обозначения: 4B, 7B, 8B, 14B, 32B, 70B.

Буква B в них означает billion, то есть миллиард параметров. Например, модель 7B содержит примерно семь миллиардов параметров, а модель 70B – около семидесяти миллиардов.

Чем больше модель, тем выше её потенциальные возможности, но тем больше требуется памяти и вычислительной мощности. Поэтому при локальном запуске важно выбирать модель, соответствующую возможностям конкретного компьютера.

II. ПРОГРАММА LLAMA.CPP КАК ИНСТРУМЕНТ ЛОКАЛЬНОГО ЗАПУСКА

Программный комплекс llama.cpp – это проект с открытым исходным кодом, предназначенный для запуска больших языковых моделей на локальных устройствах [1]. Он написан преимущественно на C и C++, что позволяет эффективно использовать ресурсы компьютера и запускать модели без сложной программной инфраструктуры.

Другими словами, llama.cpp можно рассматривать как программный «движок» для запуска языковой модели.

Сама модель хранится в отдельном файле, а llama.cpp загружает этот файл, размещает необходимые данные в памяти и выполняет вычисления.

Для понятной аналогии можно сравнить llama.cpp с медиаплеером. Видеофайл сам по себе содержит фильм, но для его просмотра нужен проигрыватель. Точно так же файл модели содержит параметры нейросети, но для работы с ним нужна программа, способная этот файл запустить.

Программный комплекс llama.cpp поддерживает:

- запуск моделей на CPU;
- ускорение на GPU;
- NVIDIA CUDA;
- Apple Metal;
- Vulkan и другие backend-решения;
- серверный режим работы;
- веб-интерфейс;
- API, совместимый с OpenAI-подобными клиентами;
- модели в формате GGUF.

Для широкой аудитории особенно важен режим llama-server. Он позволяет запустить локальный сервер, после чего пользователь может открыть браузер и взаимодействовать с моделью через обычный чат.

III. ФОРМАТ GGUF: ОПИСАНИЕ И НАЗНАЧЕНИЕ

GGUF – это специальный формат файла, предназначенный для хранения языковых моделей, совместимых с llama.cpp и рядом других программ. Файл модели обычно имеет расширение: *.gguf

GGUF можно представить как контейнер, внутри которого находятся данные, необходимые для запуска модели.

Внутри GGUF-файла хранятся числовые веса модели и служебная информация, позволяющая программе правильно загрузить и использовать модель. Логическая структура GGUF представлена на Рис. 1, рассмотрим её подробнее.

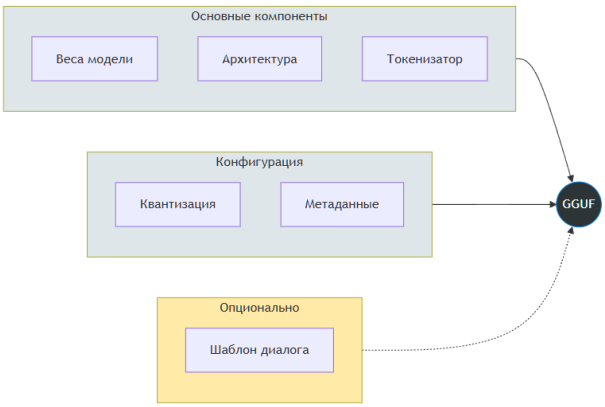


Рис. 1. Логическая структура GGUF

Основные компоненты (обязательный блок). Верхняя группа объединяет данные, без которых модель не может функционировать:

- Веса модели – числовые параметры нейронной сети, полученные в результате обучения; именно они определяют «знания» модели.
- Архитектура – описание структуры сети (тип модели, число слоёв, размерность скрытых состояний и т. п.), необходимое программе для правильной интерпретации весов.
- Токенизатор – правила и словарь, по которым текст разбивается на токены и обратно собирается из них.

Конфигурация (служебный блок). Средняя группа содержит вспомогательную информацию, определяющую способ хранения и параметры запуска модели:

- Квантизация – сведения о том, в каком формате и с какой точностью сохранены веса (например,

Q4_K_M, Q8_0), что напрямую влияет на размер файла и требования к памяти.

- Метаданные – общая справочная информация о модели: имя, версия, автор, контекстное окно, специальные токены и др.

Опционально (дополнительный блок). Нижняя группа выделена отдельным цветом и связана с контейнером пунктирной стрелкой, что подчёркивает необязательный характер её содержимого:

- Шаблон диалога (chat template) – формат оформления сообщений пользователя и ассистента; присутствует преимущественно в инструктивных и чат-моделях.

Таким образом, важно различать три понятия:

- llama.cpp – программа для запуска модели;
- GGUF – формат файла модели;
- файл *.gguf – конкретная модель, которую загружает программа.

Без файла модели llama.cpp не сможет отвечать на запросы. И наоборот, файл *.gguf сам по себе не является запускаемой программой: для работы с ним нужен llama.cpp или совместимый инструмент.

Значение GGUF состоит в том, что он упрощает распространение и запуск моделей. Пользователь часто скачивает один файл, указывает его в команде запуска и получает работающую модель. Это снижает технический порог входа по сравнению с ситуациями, когда модель состоит из множества отдельных файлов конфигурации, токенизатора и весов [2].

IV. Квантизация как способ снижения требований к ресурсам

Большие языковые модели состоят из огромного количества чисел – весов модели. Если хранить их с высокой точностью, файл модели может занимать десятки или даже сотни гигабайт. Для обычного персонального компьютера это зачастую слишком много. Для решения этой проблемы используется квантизация.

Квантизация – это метод уменьшения размера модели за счёт более компактного хранения числовых параметров. Например, вместо 16-битного или 32-битного представления чисел можно использовать 8-битное, 6-битное, 5-битное или 4-битное.

Смысл можно объяснить через аналогию со сжатием изображения. Фотографию можно сохранить в максимальном качестве, но файл будет большим. Можно применить умеренное сжатие: размер файла уменьшится, а качество для обычного просмотра останется приемлемым. Если сжатие чрезмерное, качество заметно ухудшится. С языковыми моделями ситуация похожа: умеренная квантизация снижает требования к памяти при сохранении приемлемого качества ответов.

На практике пользователь часто встречает следующие обозначения квантизации, которые представлены в Табл. I.

ТАБЛИЦА I
СРАВНИТЕЛЬНАЯ ХАРАКТЕРИСТИКА МЕТОДОВ КВАНТИЗАЦИИ GGUF-МОДЕЛЕЙ И РЕКОМЕНДАЦИИ ПО ИХ ВЫБОРУ

Метод	Общая характеристика	Когда выбирать
Q2_K	Минимальный размер, заметное снижение качества (деградация ответов).	Только если модель очень большая, а памяти критически мало.
Q3_K_M	Небольшой размер, среднее качество. Возможны логические ошибки.	Для слабых компьютеров или смартфонов.
Q4_K_M	«Золотой стандарт». Оптимальный баланс размера и интеллекта.	Лучший вариант для первого запуска и повседневного использования.
Q5_K_M	Высокое качество, близкое к оригиналу, умеренный размер.	Если RAM или VRAM позволяют (на 20-30% тяжелее Q4).
Q6_K	Очень высокое качество, минимальные потери точности.	Для мощных систем, когда важны нюансы в ответах.
Q8_0	Практически неотличимо от исходной модели (FP16).	Если памяти в избытке, и вы хотите максимум от модели.

Для большинства пользователей оптимальным первым выбором является Q4_K_M. Такая квантизация обычно обеспечивает разумный баланс между качеством, скоростью и требованиями к памяти.

V. Требования к аппаратному обеспечению

Локальный запуск языковой модели возможен на разных компьютерах, но производительность будет в значительной степени зависеть от оборудования.

Ниже представлены важнейшие характеристики компьютера, влияющие на запуск модели:

- объем ОЗУ;
- объем видеопамяти;
- производительность процессора;
- скорость работы видеокарты;
- доступное место на диске.

Модель может работать очень медленно, если она запускается только в режиме CPU. Так как требуется значительно высокие вычислительные ресурсы LLM.

Рекомендуемые языковые модели в зависимости от уровня аппаратной конфигурации представлены в Табл. II.

ТАБЛИЦА II
РЕКОМЕНДУЕМЫЕ ЯЗЫКОВЫЕ МОДЕЛИ В ЗАВИСИМОСТИ ОТ УРОВНЯ АППАРАТНОЙ КОНФИГУРАЦИИ

Уровень системы	Примерная конфигурация	Подходящие модели
Минимальный	8 ГБ RAM, без дискретной GPU	3–4B в Q4
Базовый	16 ГБ RAM	7–8B в Q4
Комфортный	16–32 ГБ RAM, GPU 8–12 ГБ VRAM	8–14B в Q4/Q5
Мощный	32–64 ГБ RAM, GPU 16–24 ГБ VRAM	27–32B в Q4
Продвинутый	64+ ГБ RAM/VRAM	крупные модели 70B и выше

Размер файла модели не равен полному объёму памяти, который понадобится при запуске. Дополнительно память расходуется на контекст, внутренние буферы и KV-кеш. Размеры моделей и требования к видеопамяти представлены в Табл. III.

ТАБЛИЦА III
РАЗМЕРЫ МОДЕЛЕЙ И ТРЕБОВАНИЯ К ВИДЕОПАМЯТИ

Размер модели	Кванти-зация	Размер файла	Требуется видео-памяти
3–4B	Q4	2,5–3,5 ГБ	4–6 ГБ
7–8B	Q4	4–6 ГБ	6–8 ГБ
12–14B	Q4	7–10 ГБ	10–14 ГБ
27–32B	Q4	16–22 ГБ	24+ ГБ
70B	Q4	40+ ГБ	48+ ГБ

Можно сформулировать практическое правило: желательно иметь запас памяти на несколько гигабайт больше, чем размер файла модели.

VI. ИСПОЛЬЗОВАНИЕ GPU И ПОДГОТОВКА NVIDIA CUDA

Для ускорения локального инференса (использование обученной модели для решения конкретной задачи) удобно использовать видеокарту. Особенно распространён вариант с видеокартами NVIDIA, поскольку llama.cpp поддерживает CUDA-сборки. CUDA (Compute Unified Device Architecture) – это программно-аппаратная архитектура параллельных вычислений, которая позволяет использовать графический процессор (GPU) для решения сложных неграфических задач [3].

Если пользователь планирует запускать модель на NVIDIA GPU, то ему необходимо установить:

- актуальный драйвер NVIDIA;
- актуальную версию CUDA Toolkit;
- CUDA-сборку llama.cpp.

Драйвер NVIDIA обеспечивает взаимодействие операционной системы с видеокартой. CUDA позволяет использовать видеокарту для вычислений общего назначения, а не только для графики. В случае LLM это особенно важно, потому что значительная часть вычислений может быть перенесена с CPU на GPU.

Основные этапы настройки видеокарты NVIDIA для запуска ИИ-моделей представлены в Табл. IV.

ТАБЛИЦА IV
ОСНОВНЫЕ ЭТАПЫ НАСТРОЙКИ ВИДЕОКАРТЫ NVIDIA

Этап	Ресурс / Команда	Описание
1. Загрузка драйверов	nvidia.com/drivers	Официальные драйверы для работы видеокарты.
2. Загрузка CUDA	developer.nvidia.com/cuda-downloads	Инструментарий для вычислений на GPU (необходим для ИИ).
3. Перезагрузка	—	Обязательное действие после установки ПО.
4. Проверка системы	nvidia-smi (в терминале/командной строке)	Команда для проверки статуса видеокарты и версий драйверов.

В Табл. V показан вывод информации при выполнении команды nvidia-smi в терминале или в powershell программах.

ТАБЛИЦА V
ОЖИДАЕМЫЕ ДАННЫЕ В ВЫВОДЕ NVIDIA-SMI

Параметр	Значение	Комментарий
Видеокарта	NVIDIA GeForce RTX 2060	8 ГБ видеопамяти
Версия драйвера (KMD)	610.47	Драйвер установлен корректно
CUDA (UMD)	13.3	Поддержка CUDA активна
Режим драйвера	WDDM	Windows Display Driver Model
Температура	37°C	Нормальная
Потребление энергии	18W / 184W	Низкая нагрузка
Использование GPU	22%	Есть активная нагрузка
Использование памяти	7902 MiB / 8192 MiB	⚠ Почти вся память занята

VII. МЕТОДИКА ЛОКАЛЬНОГО ЗАПУСКА

Практическая реализация локального запуска включает четыре основных этапа:

Подготовка оборудования и драйверов.

- Скачивание сборки llama.cpp.
- Скачивание модели в формате GGUF.
- Запуск локального сервера инференса.

Важным фактором является выбор сборки llama.cpp. Готовые сборки доступны на странице релизов [1]. Критерии выбора сборки llama.cpp в зависимости от операционной системы и аппаратной конфигурации представлены в Табл. VI.

ТАБЛИЦА VI
КРИТЕРИИ ВЫБОРА СБОРКИ LLAMA.CPP

Конфигурация	Сборка	Примечание
Windows + NVIDIA GPU	llama-bXXXX-bin-win-cuda-cu12.x-x64.zip	Самая стабильная. Работает на CUDA 12 и 13.
Windows (без видеокарты)	llama-bXXXX-bin-win-cpu-x64.zip	Только для процессора (работает медленно).
Linux (Ubuntu) + NVIDIA GPU	llama-bXXXX-bin-ubuntu-x64-cuda-cu12.x.tar.gz	Для работы через терминал Linux.
macOS (M1, M2, M3)	llama-bXXXX-bin-macos-arm64.zip	Использует встроенную графику Apple (Metal).

Вторым важным фактором является выбор модели. Модели в формате GGUF можно найти на платформе Hugging Face [4]

Готовые GGUF-квантизации часто публикуются в коллекциях различных авторов и команд [5], и позволяют в интерактивном режиме выбираем модели, помеченные как: Instruct, Chat, IT. Эти термины подразумевают, что модель обучена следовать инструкциям пользователя. Если у нас есть компьютер как минимум с 16 ГБ оперативной памяти, обычно имеет смысл для первого запуска выбрать модель 7–8B в квантовании Q4. Для более слабого компьютера лучше начать с модели 3–4B.Скачанный файл. gguf удобно поместить в папку с llama.cpp. Для простоты его можно переименовать: model.gguf

VIII. ПРИМЕР ЗАПУСКА МОДЕЛИ

Рассмотрим типовой сценарий для компьютера с Windows 10/11, видеокартой NVIDIA и 8 ГБ оперативной памяти.

A. Запуск сервера (GPU-режим)

После компиляции или загрузки бинарных файлов llama.cpp выполняется переход в рабочую директорию в PowerShell:

```
cd C:\llama-cpp
```

Запуск сервера осуществляется следующей командой:

```
.\llama-server.exe -m
C:\llama.cpp\model.gguf -c 10200 --host
127.0.0.1 --port 8080 -ngl 999 -t 12 -fa on
-b 512 -ub 512 -np 1 --cache-ram 0 --
reasoning off
```

Основные параметры командной строки для запуска сервера llama.cpp и их назначение приедены в Табл. VII.

ТАБЛИЦА VII
ОСНОВНЫЕ ПАРАМЕТРЫ ЗАПУСКА LLAMA.CPP

Параметр	Назначение
-m	путь к модели в формате GGUF
-c 10200	размер контекстного окна (10 200 токенов)
--host 127.0.0.1	ограничение доступа локальным интерфейсом
--port 8080	порт HTTP-сервера
-ngl 999	попытка загрузки всех слоёв модели в видеопамять GPU
-t 12	количество потоков CPU
-fa on	включение Flash Attention
-b 512	размер batch для обработки токенов
-ub 512	размер micro-batch (внутренняя разбивка вычислений)
-np 1	количество параллельных последовательностей (один пользовательский поток)
--cache-ram 0	отключение размещения KV-кэша в оперативной памяти (используется VRAM)
--reasoning off	отключение режима расширенного рассуждения (если поддерживается сборкой)

B. Конфигурационные особенности

Используемая модель (model.gguf) загружается в формате GGUF, предназначенном для эффективного локального инференса в среде llama.cpp. Конкретные характеристики модели (число параметров, тип архитектуры, уровень квантования) определяют требования к объёму оперативной и видеопамяти, а также итоговую производительность генерации.

Параметр -ngl 999 задаёт попытку полной выгрузки всех слоёв модели на графический ускоритель. Фактическое количество слоёв, размещённых в видеопамяти, автоматически ограничивается доступным объёмом VRAM. В случае нехватки видеопамяти значение параметра -ngl должно быть уменьшено для обеспечения стабильной работы.

Размер контекстного окна 10200 токенов приводит к увеличению объёма KV кэша, что напрямую влияет на потребление видеопамяти и оперативной памяти. Требуемый объём ресурсов зависит от архитектуры модели и её размерности; для моделей среднего и крупного масштаба может потребоваться GPU с увеличенным объёмом VRAM.

Параметры -b 512 (batch size) и -ub 512 (micro-batch size) определяют объём токенов, обрабатываемых за один вычислительный шаг. Увеличение этих значений повышает пропускную способность инференса при достаточном объёме памяти, однако может привести к ошибкам выделения ресурсов на менее производительном оборудовании.

После успешного запуска сервер будет доступен в браузере по адресу:

```
http://127.0.0.1:8080
```

Если видеокарты нет, используется CPU-вариант:

```
.\llama-server.exe -m model.gguf -c 4096 -
-host 127.0.0.1 --port 8080 -ngl 0 -t 6 --
temp 0.7
```

Если видеокарта есть, но модель не помещается полностью в видеопамять, можно уменьшить число слоёв, загружаемых на GPU:

```
.\llama-server.exe -m model.gguf -c 4096 --host 127.0.0.1 --port 8080 -ngl 20 -t 6 -fa --temp 0.7
```

Для проверки работы модели достаточно открыть браузер и ввести простой запрос, например: «Напиши код формата html веб страницы с кнопками и ссылками»

Если модель отвечает осмысленно, значит программа, модель и веб-интерфейс работают корректно.

IX. РАБОТА ЧЕРЕЗ API

llama-server может использоваться не только через браузер, но и как локальный API-сервер. Это делает его полезным для разработчиков, образовательных учреждений и организаций, которые хотят встроить локальную модель в собственные программные решения.

Пример запроса через curl для Windows:

```
curl http://127.0.0.1:8080/v1/chat/completions ^
    -H "Content-Type: application/json" ^
    -d "{\"messages\":[{\"role\":\"user\",\"content\":\"Объясни простыми словами, что такое инференс нейросети\"}],\"temperature\":0.7}"
```

Пример для Linux и macOS:

```
curl http://127.0.0.1:8080/v1/chat/completions \
    -H "Content-Type: application/json" \
    -d '{
      "messages": [
        {
          "role": "user",
          "content": "Объясни простыми словами, что такое инференс нейросети"
        }
      ],
      "temperature": 0.7
    }'
```

Пример на Python:

```
from openai import OpenAI

client = OpenAI(
    base_url="http://127.0.0.1:8080/v1",
    api_key="not-needed"
)

response = client.chat.completions.create(
    model="local-model",
    messages=[
        {
            "role": "user",
            "content": "Объясни, что означает запуск модели на локальном компьютере"
        }
    ],
    temperature=0.7
)

print(response.choices[0].message.content)
```

Локальная модель может использоваться не только как чат, но и как компонент программной системы.

X. НАСТРОЙКА ПРОИЗВОДИТЕЛЬНОСТИ

Производительность локальной модели зависит от нескольких факторов: размера модели, типа квантизации, объёма оперативной памяти, видеопамяти, мощности процессора и параметров запуска.

Рекомендуемые значения параметров llama.cpp для настройки контекста, вычислений и стиля генерации ответов приведены в Табл. VIII.

ТАБЛИЦА VIII
РЕКОМЕНДУЕМЫЕ ЗНАЧЕНИЯ ПАРАМЕТРОВ LLAMA.CPP

Параметр	Значение
-c 4096	обычный диалог, меньший расход памяти
-c 8192	более длинный контекст
-ngl 999	попытка загрузить максимум слоёв модели на GPU
-ngl 0	запуск только на CPU
--temp 0.3	более строгие и предсказуемые ответы
--temp 0.7	баланс точности и разнообразия
--temp 0.9	более свободная генерация

Размер контекста определяет, какой объем информации нейросеть может обработать за один раз. Если контекст большой, модель лучше помнит детали долгого разговора, но это требует больше компьютерной памяти. Для обычного общения вполне хватает 4096 или 8192 токенов. Огромные значения нужны в основном для работы с очень длинными документами или целыми книгами.

Другой важный параметр – это температура. Она влияет на то, насколько случайным или предсказуемым будет результат. Нейросеть не просто выдает готовое слово, а выбирает из списка возможных вариантов с разной вероятностью. Температура как раз регулирует этот выбор. Само название пришло из физики: чем выше температура, тем хаотичнее ведут себя частицы. В случае с текстом логика похожая.

При низких значениях, например 0,3, модель выбирает самые очевидные и логичные варианты, поэтому ответы получаются стабильными и строгими. Если же выставить 0,9 или выше, нейросеть начнет чаще выбирать редкие слова. Текст станет более живым и оригинальным, но в нем могут появиться фактические ошибки. Значение 0,7 обычно считается

золотой серединой и подходит для большинства повседневных задач.

XI. ОБСУЖДЕНИЕ

Локальный запуск больших языковых моделей показывает, что современные ИИ-инструменты постепенно переходят из категории облачных сервисов в категорию персонального программного обеспечения. Это имеет важные последствия для образования, исследований, малого бизнеса и индивидуального использования.

Главное преимущество локального подхода состоит в повышении контроля над данными. Если модель работает на компьютере пользователя, запросы не передаются внешнему сервису. Это особенно важно при работе с внутренними документами, учебными материалами, исследовательскими заметками и персональной информацией.

Второе преимущество – автономность. После загрузки модели пользователь может работать без постоянного подключения к интернету. Это полезно в условиях ограниченной связи, закрытых сетей или необходимости воспроизводимой работы в локальной среде.

Третье преимущество – гибкость. Пользователь может выбирать модель, квантизацию, размер контекста, параметры генерации и способ интеграции. В облачных сервисах эти параметры часто скрыты или ограничены.

Однако локальный запуск не является полной заменой облачным системам. Крупные облачные модели могут обладать более высоким качеством, большим контекстом и лучшей инфраструктурой. Кроме того, локальные модели требуют подходящего оборудования и определённой технической подготовки.

Особое значение имеет формат GGUF. Он снижает сложность распространения моделей и делает возможным сценарий, при котором пользователь скачивает один файл модели и запускает его через llama.cpp. В сочетании с квантизацией это позволяет использовать LLM на обычных компьютерах, а не только на специализированных серверах.

ЗАКЛЮЧЕНИЕ

Сейчас запустить нейросеть у себя на компьютере стало намного проще. Появились удобные открытые инструменты вроде llama.cpp и формат GGUF, а технологии сжатия позволяют работать с моделями даже на типичной конфигурации пользовательского компьютера. Раньше для этого обязательно нужны были облачные сервисы, а теперь всё можно сделать локально.

Формат GGUF сильно упростил процесс локального запуска больших языковых моделей. По сути, это один файл, в котором уже есть всё необходимое: и сами данные модели, и её настройки. На практике всё выглядит довольно прямолинейно: необходимо скачать программу, файл с моделью, запустить его через команду и открыть интерфейс в обычном браузере.

Конечно, придется немного подготовиться: проверить комплектующие, установить драйверы для видеокарты и разобраться с запуском сервера. Зато такой подход дает полную приватность, так как пользовательские данные не передаются в Интернет. Кроме того, пользователь не зависит от сторонних сервисов и можете настраивать модель под свои нужды или встраивать её в свои проекты.

Основное ограничение здесь – это производительность компьютера. Не всякая конфигурация компьютера сможет успешно работать со сложными моделями с хорошим качеством ответов. Пока можно прогнозировать, что локальные нейросети вряд ли полностью вытеснят облака, но они стали отличной альтернативой для тех, кто хочет сам управлять искусственным интеллектом на своем устройстве.

ЛИТЕРАТУРА

[1] ggml-org. llama.cpp : репозиторий исходного кода / ggml-org. — Текст: электронный // GitHub [сайт]. — URL: https://github.com/ggml-org/llama.cpp (дата обращения: 01.07.2026).

[2] Chia, A. GGUF Format: A Complete Guide to Local LLM Inference / A. Chia. — Текст : электронный // DataCamp : [сайт]. — 2024. — URL: https://www.datacamp.com/tutorial/gguf-format-a-complete-guide (дата обращения: 28.06.2026).

[3] Васильянов, А. И. Устройства NVIDIA серии Jetson для работы с нейросетями / А. И. Васильянов, В. И. Бондаренко. — Текст : электронный // Вестник Донецкого национального университета. Серия Г: Технические науки. — 2025. — № 1. — С. 93–103. — ISSN 2663-4228. — URL: https://e.lanbook.com/journal/issue/371619 (дата обращения: 01.07.2026).

[4] Hugging Face : [сайт]. — Текст : электронный. — URL: https://huggingface.co (дата обращения: 01.07.2026).

[5] Коллекции квантизованных моделей GGUF от Unsloth. — Текст : электронный // Hugging Face : [сайт]. — URL: https://huggingface.co/unsloth (дата обращения: 01.07.2026).

Информация об авторах

Мартынов Сергей Владимирович – магистрант направления подготовки «Инфокоммуникационные технологии и системы связи» Астраханского государственного технического университета, г. Астрахань, Россия, sergeytmm@mail.ru, ORCID: 0009-0008-4739-0719

Осовский Алексей Викторович – кандидат технических наук, доцент кафедры «Связь» Астраханского государственного технического университета, г. Астрахань, Россия, a_osovskiy@mail.ru, ORCID: 0000-0002-3174-9765

Кутузов Денис Валерьевич – кандидат технических наук, доцент кафедры «Связь» Астраханского государственного технического университета, г. Астрахань, Россия, d_kutuzov@mail.ru, ORCID: 0000-0001-7963-1259

Experience deploying large language models on local computing systems

Sergey Martynov, Alexey Osovsky, Denis Kutuzov

Astrakhan State Technical University, Astrakhan, Russia

***Abstract* – This article examines the practical aspects of local deployment of Large Language Models (LLM) on personal computing systems. It provides an overview of the local inference architecture, the llama.cpp software suite, and the GGUF model storage format, which ensures the efficient use of modern language models on user hardware. The principles of model quantization and their impact on RAM and video memory requirements are considered, and recommendations for choosing the quantization level depending on the computer hardware configuration are provided. The article describes the software and hardware requirements, the specifics of using NVIDIA graphics accelerators with CUDA technology, and a methodology for preparing an environment for local model execution. Practical recommendations are presented for choosing llama.cpp builds, loading models, configuring launch parameters, organizing a local inference server, and using the OpenAI-compatible API for integrating language models into application software systems. The advantages of the local approach related to data confidentiality, autonomous operation, and flexibility of model configuration are discussed, as well as its limitations due to the computing resources of personal computers. The results obtained can be used by developers, researchers, educators, and specialists interested in deploying personal artificial intelligence systems without the use of cloud infrastructure.**

***Keywords* – large language models, personal artificial intelligence, local inference, llama.cpp, GGUF, quantization, CUDA**

REFERENCES

[1] ggml-org. llama.cpp : repozitorij iskhodnogo koda / ggml-org. Tekst : elektronnyj // GitHub : [sajt]. — URL: https://github.com/ggml-org/llama.cpp (data obrashcheniya: 01.07.2026).

[2] Chia, A. GGUF Format: A Complete Guide to Local LLM Inference / A. Chia. — Tekst : elektronnyj // DataCamp : [sajt]. — 2024. — URL: https://www.datacamp.com/tutorial/gguf-format-a-complete-guide (data obrashcheniya: 28.06.2026).

[3] Vasil'yanov, A. I. Ustrojstva NVIDIA serii Jetson dlya raboty s nejrosetyami / A. I. Vasil'yanov, V. I. Bondarenko. — Tekst : elektronnyj // Vestnik Doneckogo nacional'nogo universiteta. Seriya G: Tekhnicheskie nauki. — 2025. — № 1. — S. 93–103. — ISSN 2663-4228. — URL: https://e.lanbook.com/journal/issue/371619 (data obrashcheniya: 01.07.2026).

[4] Hugging Face : [sajt]. — Tekst : elektronnyj. — URL: https://huggingface.co (data obrashcheniya: 01.07.2026).

[5] Kollekcii kvantizovannyh modelej GGUF ot Unsloth. — Tekst : elektronnyj // Hugging Face : [sajt]. — URL: https://huggingface.co/unsloth (data obrashcheniya: 01.07.2026).

Information about the authors

Sergey V. Martynov, master's student in the program "Infocommunication Technologies and Communication Systems" of the Astrakhan State Technical University, Astrakhan, Russia, sergeytmm@mail.ru, ORCID: 0009-0008-4739-0719

Alexey V. Osovsky, Cand. of Techn. Sci., Associate professor in the Communications Department at Astrakhan State Technical University, Astrakhan, Russia, a_osovskiy@mail.ru, ORCID: 0000-0002-3174-9765

Denis V. Kutuzov, Cand. of Techn. Sci., Associate professor, Department of Communications, Astrakhan State Technical University, Astrakhan, Russia, d_kutuzov@mail.ru, ORCID: 0000-0001-7963-1259

Проактивное управление инфраструктурой многоквартирных домов: обзор современных цифровых платформ

Ф.С. Таргачева, А.Р. Мамлеева

Астраханский государственный технический университет, Астрахань, Россия

Аннотация **– В статье рассматривается текущее состояние управления многоквартирными домами в условиях цифровизации жилищно-коммунального хозяйства. Отмечается, что диспетчерские службы управляющих организаций остаются ключевым элементом операционного управления, однако в большинстве случаев продолжают функционировать в рамках реактивной модели, ориентированной на устранение последствий уже возникших инцидентов. Проанализированы основные проблемы эксплуатации жилого фонда, включая высокие внеплановые затраты, рост числа обращений жителей, простои инженерных систем и юридические риски. Выполнен обзор распространенных диспетчерских платформ, применяемых в сфере жилищно-коммунального хозяйства, с выделением их функциональных возможностей и ограничений. Особое внимание уделено критериям эффективности диспетчерских систем и роли аналитических инструментов. Обоснована необходимость формирования современного рабочего места диспетчера, ориентированного на проактивное управление, интеграцию данных и интеллектуальную поддержку принятия решений.**

Ключевые слова **– диспетчерские системы, управление многоквартирными домами, проактивное управление, жилищно-коммунальное хозяйство, аварийность в ЖКХ, цифровые платформы ЖКХ, аналитика данных.**

ВВЕДЕНИЕ

Современный жилищный фонд крупных городов отличается высокой концентрацией жилых объектов и значительной долей зданий с высоким уровнем физического износа. По данным официальной статистики, более 70 % населения Российской Федерации проживает в многоквартирных домах, при этом около 45 % таких зданий эксплуатируются свыше тридцати лет [1]. В этих условиях устойчивость систем водоснабжения, отопления, электроснабжения и лифтового хозяйства напрямую зависит от качества оперативного управления.

Центральное место в этой системе занимает диспетчерская служба управляющей организации. Именно через нее проходят обращения жителей, сигналы аварийных служб, контроль сроков выполнения работ и формирование отчетности для государственных информационных систем. Несмотря на активное внедрение цифровых решений в жилищно-коммунальном хозяйстве (ЖКХ), на практике диспетчерская деятельность часто строится на разрозненных инструментах и ручных операциях. Это приводит к увеличению времени реакции, дублированию информации и росту операционных затрат.

Дополнительное давление формируется нормативными изменениями. В последние годы усиливается ответственность управляющих организаций за несоблюдение сроков устранения аварий, а использование государственных цифровых платформ, в том числе ГИС ЖКХ, становится обязательным [2, 3]. В совокупности эти факторы формируют запрос на пересмотр роли диспетчера и переход от реактивной модели управления к проактивной, ориентированной на предупреждение проблем.

Цель данного обзора – систематизировать существующие проблемы в управлении многоквартирными домами (МКД), проанализировать недостатки современных диспетчерских систем и обосновать необходимость перехода к проактивному управлению на основе цифровых платформ, интегрирующих аналитику данных и машинное обучение.

I. СОВРЕМЕННЫЕ ПРОБЛЕМЫ УПРАВЛЕНИЯ МНОГОКВАРТИРНЫМИ ДОМАМИ

Современная система управления многоквартирными домами характеризуется совокупностью взаимосвязанных проблем, затрагивающих технические, экономические, организационные и социальные аспекты. Одной из наиболее острых остается проблема износа инженерных сетей. По оценкам экспертов, в ряде регионов доля внутридомовых инженерных систем с износом свыше 60 % превышает 45 % [4]. Это приводит к росту числа аварий, особенно в отопительный период, и

увеличению нагрузки на аварийно-диспетчерские службы.

Финансовые потери управляющих организаций во многом обусловлены реактивным характером управления. Экстренные ремонты, выполняемые в сжатые сроки, как правило, обходятся на 20–30 % дороже плановых профилактических мероприятий [5]. При этом отсутствие достоверной аналитики по повторяемости инцидентов не позволяет формировать обоснованные программы капитального и текущего ремонта. В результате значительная часть средств расходуется на устранение последствий, а не на предотвращение причин.

Отдельную группу проблем формируют правовые и репутационные риски. Количество жалоб собственников в жилищные инспекции и органы прокуратуры стабильно растет, что подтверждается статистикой обращений в ГИС ЖКХ [6]. Судебные иски, связанные с ненадлежащим содержанием общего имущества, все чаще заканчиваются взысканием штрафов и компенсаций. По оценкам профильных ассоциаций, до 10–12 % обращений жителей перерастают в формальные претензии, требующие юридического сопровождения.

Организационные сложности во многом связаны с перегрузкой диспетчерского персонала. В условиях круглосуточного режима работы диспетчеры вынуждены одновременно принимать звонки, регистрировать заявки, контролировать исполнение и формировать отчетность. Отсутствие автоматизированных подсказок и приоритизации заявок приводит к субъективным решениям и ошибкам. На практике это проявляется в том, что часть критически важных инцидентов обрабатывается с задержкой, тогда как менее значимые обращения получают непропорционально большое внимание.

Отдельного внимания требует фрагментация информационных потоков. В УО часто используются разрозненные программные продукты: отдельные системы для учета заявок, бухгалтерии, взаимодействия с подрядчиками и выгрузки данных в ГИС ЖКХ [7]. Отсутствие сквозной интеграции приводит к дублированию информации и снижению ее достоверности. По оценкам экспертов, до 15 % записей в локальных базах данных содержат расхождения с официальной отчетностью [8].

Структура ключевых проблем управления МКД обобщена на Рис. 1. На графике видно, что доминирующей угрозой для УК остаются финансовые издержки (суммарно до 40% с учетом судебных исков). Эти расходы приводят к внеплановых затратам управляющих организаций (УО). Такие непредвиденные расходы «съедают» весь бюджет компаний, и они массово банкротятся. Из-за этого за последний год количество закрывающихся и разоряющихся УК выросло почти в 3 раза. Компании физически не справляются с судами и долгами, поэтому им жизненно необходимы цифровые инструменты, которые могут помочь снизить эти риски.

Рис. 1. Доли основных проблем управления многоквартирными домами

Таким образом, современные проблемы управления МКД имеют комплексный характер и требуют системных изменений, затрагивающих в том числе подходы к организации диспетчерской деятельности.

II. ОСНОВЫ ПРОАКТИВНОГО УПРАВЛЕНИЯ В ЖКХ

Проактивное управление – это стратегический подход, направленный на предупреждение проблем, а не на устранение их последствий. В его основе лежит непрерывный цикл: сбор данных, анализ, прогноз и принятие мер [9].

Переход от реактивной модели к проактивной в ЖКХ невозможен без использования современных технологий анализа данных. Например, методы машинного обучения, такие как LightGBM и Random Forest, позволяют на основе собранных данных предсказывать вероятность поломок оборудования – лифтов или насосов в ближайшие 30 дней, что помогает планировать профилактические ремонты и снижать число аварий на 25–30 % [10].

Для обнаружения резких отклонений в работе систем применяются специальные алгоритмы и статистические методы, которые выявляют отклонения – например, резкий рост заявок из одного дома или увеличение времени их решения [11, 12].

Кроме того, важную роль играют ETL-процессы (Extract, Transform, Load), которые извлекают, преобразуют и загружают данные из разных источников – таких как ГИС ЖКХ, базы заявок и показания датчиков для дальнейшего анализа.

В итоге эти технологии позволяют переходить от реагирования на жалобы к управлению, основанному на прогнозах, и проводить техническое обслуживание еще до появления серьезных проблем [13].

III. ОБЗОР СУЩЕСТВУЮЩИХ ДИСПЕЧЕРСКИХ СИСТЕМ

Развитие цифровых инструментов в сфере управления многоквартирными домами привело к формированию нескольких устойчивых классов диспетчерских решений. Эти системы различаются по архитектуре, функциональной направленности и роли в управленческих процессах. Ниже рассмотрены основные группы решений, получившие распространение в практике управляющих организаций.

A. Классические диспетчерские решения

Наибольшее распространение получили классические диспетчерские системы, ориентированные прежде всего на организацию аварийно-диспетчерского обслуживания и прием обращений жителей. Их ключевая задача заключается в обеспечении непрерывного канала связи между жильцами и эксплуатационными службами, а также в контроле сроков устранения неисправностей. К ним относится платформа «Диспетчер 24», широко используемая в крупных УО и аутсорсинговых диспетчерских центрах [14]. Основной функционал системы включает прием звонков, регистрацию заявок, распределение задач между исполнителями и контроль сроков выполнения. Преимуществом является масштабируемость и возможность работы в режиме 24/7, однако аналитические функции ограничены базовой статистикой.

Помимо ранее рассмотренной платформы «Диспетчер 24», к данной категории относятся системы «АДС ЖКХ» [15] и «ЕДДС ЖКХ» [16], используемые как в коммерческих УО, так и в муниципальных структурах. Решение «АДС ЖКХ» ориентировано на централизованный учет аварийных заявок, регистрацию обращений по телефону и формирование регламентированной отчетности для надзорных органов. Система «ЕДДС ЖКХ», как правило, интегрируется с муниципальными экстренными службами и используется для обработки инцидентов повышенной критичности.

Общей особенностью классических решений является их высокая устойчивость к нагрузкам и соответствие нормативным требованиям аварийно-диспетчерского обслуживания. Вместе с тем функциональность таких систем ограничивается фиксацией фактов и контролем исполнения. Аналитические инструменты, как правило, сводятся к базовой статистике, а приоритизация заявок осуществляется вручную.

Классические диспетчерские системы эффективно решают задачи оперативного реагирования, однако не позволяют снижать долю внеплановых затрат и простоев инженерных систем. В структуре проблем управления МКД на их долю в наибольшей степени приходится влияние финансовых потерь и перегрузки персонала (Рис. 2).

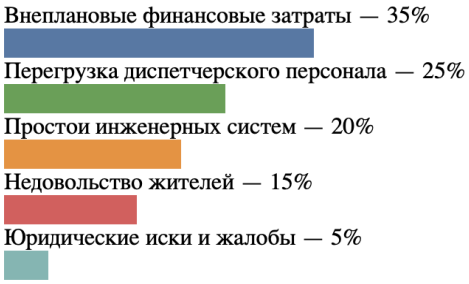


Рис. 2. Структура проблем управления МКД при использовании классических диспетчерских решений

B. Корпоративные платформы управления ЖКХ

Отдельную категорию образуют корпоративные системы управления, в которых диспетчерский функционал является частью более широкой платформы. Яркий пример – «1С: Управление ЖКХ», связанная с расчетами, бухгалтерией и контрактной работой [17]. Использование единой экосистемы снижает количество дублируемых операций и упрощает формирование отчетности. В то же время диспетчерский модуль в таких системах зачастую ориентирован на учет фактов, а не на поддержку оперативных решений.

АСУ «Жилищный стандарт» и аналогичные комплексы предлагают расширенные возможности по учету технического состояния объектов и планированию работ [18]. Однако на практике их внедрение требует значительных организационных изменений и обучения персонала, что ограничивает распространение среди небольших управляющих компаний.

Помимо системы «1С: Управление ЖКХ» и АСУ «Жилищный стандарт», к данному классу относится платформа «Парус–ЖКХ» [19], применяемая в ряде регионов для автоматизации деятельности крупных управляющих компаний и ресурсоснабжающих организаций.

Платформа «Парус–ЖКХ» обеспечивает сквозной учет процессов управления жилым фондом, включая ведение технической документации, планирование работ и финансовый контроль. Диспетчерский модуль в таких системах тесно связан с управлением ресурсами и подрядчиками, что повышает прозрачность исполнения заявок.

При этом корпоративные платформы часто ориентированы на регламентные процессы и отчетность. В условиях высокой динамики аварийных ситуаций это приводит к снижению гибкости и увеличению времени реакции. Кроме того, внедрение таких систем требует значительных организационных изменений и адаптации персонала.

Корпоративные платформы способствуют снижению финансовых и юридических рисков за счет интеграции

данных, однако не обеспечивают достаточной поддержки проактивного управления. Основные проблемы МКД в данном случае смещаются в сторону временных задержек и организационной инерции, что отражается на диаграмме для корпоративных решений (Рис. 3).

Рис. 3. Структура проблем управления МКД при использовании корпоративных платформ ЖКХ

C. Платформы «умного» и цифрового ЖКХ

С развитием концепции «умного города» на рынке появились решения, ориентированные на автоматизированный мониторинг инженерных систем и сбор телеметрических данных. Платформа «Умное ЖКХ» демонстрирует возможности автоматического контроля параметров потребления ресурсов и раннего выявления отклонений [20]. В ряде пилотных проектов внедрение таких систем позволило сократить количество аварий на 15–18 % в течение двух лет эксплуатации.

Кроме платформы «Умное ЖКХ», к данному классу относится решение «Цифровой дом» [21], внедряемое в рамках проектов «умного города» и использующее данные с приборов учета, датчиков давления, температуры и утечек.

Подобные системы позволяют выявлять отклонения в работе инженерных сетей на ранних стадиях и формировать сигналы о потенциальных неисправностях. В ряде пилотных проектов применение таких платформ привело к снижению аварийности и более равномерному распределению нагрузки на эксплуатационные службы.

Ограничением данного класса решений остается слабая интеграция с диспетчерскими процессами. Диспетчер получает значительный объем данных, однако без развитых аналитических инструментов и рекомендаций по действиям это увеличивает когнитивную нагрузку и не всегда приводит к снижению числа инцидентов.

Платформы цифрового ЖКХ позволяют снижать долю аварий и простоев инженерных систем, но без интеграции с диспетчерским рабочим местом не решают проблему перегрузки персонала. В структуре проблем МКД для данного класса решений доминируют организационные и информационные факторы (Рис. 4).

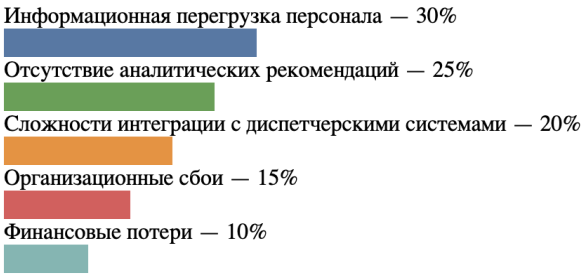


Рис. 4. Структура проблем управления МКД при использовании платформ «умного» и цифрового ЖКХ

D. Сервисы взаимодействия с жителями

Сервисы взаимодействия с жителями формируют отдельный класс решений, ориентированных на коммуникацию и повышение прозрачности деятельности управляющих организаций. Это «РосКвартал» [22], обеспечивающий прием заявок через мобильные приложения, информирование о работах и публикацию отчетов; платформы «Домопульт» [23] и «ЖКХ Онлайн» [24], предоставляющие мобильные приложения для подачи заявок, получения уведомлений и доступа к информации о начислениях и работах.

Использование таких сервисов снижает нагрузку на телефонные линии диспетчерских служб и повышает вовлеченность жителей. Вместе с тем данные решения редко используются как полноценные диспетчерские системы и, как правило, требуют интеграции с внутренними платформами управляющих организаций.

Отсутствие сквозного контроля исполнения заявок и аналитики приводит к тому, что сервисы взаимодействия решают коммуникационные задачи, но не влияют напрямую на эксплуатационную надежность жилого фонда.

Сервисы взаимодействия с жителями способствуют снижению репутационных рисков и уровня недовольства собственников, однако практически не влияют на финансовые потери и аварийность. Их вклад в структуру проблем управления МКД преимущественно связан с коммуникацией (Рис. 5).

Рис. 5. Структура проблем управления МКД при использовании сервисов взаимодействия с жителями

E. Критерии эффективности диспетчерских систем

Профессиональный стандарт 16.018 «Специалист по управлению многоквартирным домом» [25] задает основные обязанности диспетчерской службы: прием и регистрацию заявок, их распределение по приоритетам, назначение исполнителей и контроль выполнения. Однако для оценки эффективности цифровых систем диспетчеризации требуется более широкий набор критериев [26].

Во-первых, система должна легко интегрироваться с другими сервисами, такими как ГИС ЖКХ, CRM, бухгалтерские программы и платформы Интернета вещей (IoT). Во-вторых, важен удобный и понятный интерфейс с быстрой работой, а также наличие мобильной версии и поддержки push-уведомлений для диспетчеров и жителей. Также необходима аналитика: система должна показывать ключевые показатели – общее число заявок, среднее время их решения (MTTR), соблюдение стандартов обслуживания (SLA) не менее 95 %, а также количество заявок на 1000 м² управляемой площади. Важна поддержка проактивных функций, таких как прогнозирование проблем с помощью машинного обучения. Согласно международному опыту, эффективная система диспетчеризации обеспечивает среднее время решения заявки менее четырех часов и уровень удовлетворенности жителей выше 85 % [27]. Однако, помимо этих критериев, важно учитывать также гибкость системы, ее способность адаптироваться к изменяющимся требованиям и масштабироваться в зависимости от роста обслуживаемого жилищного фонда.

F. Сравнительный анализ аналогов

Сводное сравнение функциональных возможностей наиболее распространенных платформ представлено в Табл. I.

ТАБЛИЦА I
СРАВНЕНИЕ ДИСПЕТЧЕРСКИХ СИСТЕМ ЖКХ

Характеристика/ Платформа	**Классические системы**	**Корпоративные платформы**	**Платформы цифрового и «умного» ЖКХ**	**Сервисы взаимодействия с жителями**
Централизованный учет заявок	да	да	да	частично
Контроль сроков исполнения	нет	да	частично	нет
Интеграция с инженерными системами	да	частично	да	нет
Поддержка аналитики и отчетности	нет	да	частично	нет
Прогнозирование аварийных ситуаций	нет	нет	частично	частично
Поддержка проактивного управления	нет	нет	частично	частично
Масштабируемость на несколько МКД	нет	да	да	частично
Автоматизация коммуникаций с жителями	нет	нет	частично	да

Сравнение показало, что ни один из рассмотренных классов решений не обеспечивает комплексного охвата ключевых задач управления МКД. Классические диспетчерские системы ориентированы преимущественно на реактивное устранение аварий, корпоративные платформы – на учет и регламентацию процессов, цифровые платформы – на сбор данных без достаточной аналитической поддержки, а сервисы взаимодействия с жителями – на коммуникации без управленческого контроля [28]. Данный разрыв функциональности, наглядно представленный в Табл. I, формирует предпосылки для перехода к специализированному рабочему месту диспетчера, ориентированному на проактивное управление и поддержку принятия решений.

IV. КОНЦЕПЦИЯ ПРОАКТИВНОГО РАБОЧЕГО МЕСТА ДИСПЕТЧЕРА

Переход к проактивному управлению предполагает качественное изменение функциональной роли диспетчера. Современное рабочее место должно обеспечивать агрегирование данных из различных источников, визуализацию ключевых показателей и интеллектуальную поддержку принятия решений.

Использование методов машинного обучения и промышленного Интернета вещей позволяет выявлять скрытые закономерности в потоках заявок и прогнозировать вероятность отказов инженерных систем [29, 30]. В ряде отраслей применение предиктивной аналитики приводит к снижению

аварийности на 20–30 % уже в течение первого года эксплуатации [31, 32]. Для ЖКХ это означает возможность перехода от устранения последствий к планированию профилактических мероприятий.

Рабочее место диспетчера в проактивной модели представляет собой единый интерфейс, объединяющий текущие инциденты, показатели SLA, прогнозные оценки и рекомендации по перераспределению ресурсов. Это снижает нагрузку на персонал и повышает устойчивость управленческих решений.

Сопоставление существующих диспетчерских систем с требованиями проактивного управления показывает наличие значительного разрыва между текущей практикой и потенциальными возможностями цифровых технологий. В ряде случаев внедрение информационных систем ограничивается формальной автоматизацией учета заявок без изменения логики управления.

Международный опыт подтверждает, что интеграция аналитических и прогнозных инструментов в процессы управления городской инфраструктурой способствует снижению издержек и повышению надежности эксплуатации. Для сферы ЖКХ данный подход остается перспективным, но недостаточно реализованным.

ВЫВОДЫ И ЗАКЛЮЧЕНИЕ

Анализ показывает, что нужно создать современное рабочее место диспетчера для многоквартирных домов. Существующие системы в основном решают учетные задачи и работают по реактивному принципу, что неэффективно из-за роста требований и ожиданий жителей. Главные проблемы – задержки в обработке заявок, разрозненные данные и отсутствие инструментов для предупреждения аварий.

Для улучшения нужны интеллектуальные инструменты на базе машинного обучения, которые прогнозируют риски и автоматически запускают превентивные проверки. Важна удобная панель с анализом и рекомендациями, гибкая архитектура для интеграции с внешними системами и автоматизация рутинных задач, чтобы диспетчеры могли сосредоточиться на важных вопросах.

Такой подход превратит диспетчера из оператора в ключевого управленца, повышая надежность работы и комфортность проживания.

ЛИТЕРАТУРА

[1] Росстат. Жилищный фонд Российской Федерации: статистический сборник. – М.: Росстат, 2023. – 150 с.

[2] Федеральный закон от 21.07.2014 № 209-ФЗ "О государственной информационной системе жилищно-коммунального хозяйства" [Электронный ресурс]. – Режим доступа: http://www.consultant.ru/document/cons_doc_LAW_165820 (дата обращения: 12.11.2025).

[3] Федеральный закон от 13.12.2024 № 493-ФЗ "О внесении изменений в Жилищный кодекс Российской Федерации". – [Электронный ресурс]. – Режим доступа: https://www.consultant.ru/document/cons_doc_LAW_493074 (дата обращения: 12.11.2025).

[4] Концепция цифровизации городского хозяйства до 2030 года (утв. Правительством РФ) [Электронный ресурс]. – Режим доступа: http://static.government.ru/media/files/7gWr5l5t6bTctkfqRt1f64CXhWlZv6dA.pdf (дата обращения: 12.11.2025).

[5] Стратегия развития строительной отрасли и жилищно-коммунального хозяйства Российской Федерации на период до 2030 года [Электронный ресурс]. – Режим доступа: http://static.government.ru/media/files/8gWr5l5t6bTctkfqRt1f64CXhWlZv6dA.pdf (дата обращения: 12.11.2025).

[6] Прокуроры выявили более 424500 нарушений законов в сфере ЖКХ в 2024 году. – URL: https://gkhrazvitie.ru/tape/news/2025/03/prokurory-vyyavili-bolee-424-500-narushenii-zakonov-v-sfere-zhkh-v-2024-godu/ (дата обращения: 19.12.2025).

[7] Диспетчер 24: автоматизация диспетчерской службы для ЖКХ [Электронный ресурс]. – Режим доступа: https://ds24.ru/ (дата обращения: 12.11.2025).

[8] Системы автоматизации 1С ЖКХ, ТСЖ, ЖСК [Электронный ресурс]. – Режим доступа: https://www.1cbit.ru/1s-otrasli/avtomatizaciya-zhkkh-tszh/ (дата обращения: 12.11.2025).

[9] Автоматизированная система управления Жилищный стандарт [Электронный ресурс]. – Режим доступа: https://catalog.arppsoft.ru/product/6044568 (дата обращения: 12.11.2025).

[10] Проактивное обслуживание в сфере ЖКХ как тренд технологического развития. [Электронный ресурс]. – Режим доступа: https://cyberleninka.ru/article/n/proaktivnoe-obsluzhivanie-v-sfere-zhkh-kak-trend-tehnologicheskogo-razvitiya (дата обращения: 09.12.2025).

[11] Информационные системы и технологии: монография / О.И. Бабина, Н.Ю. Дюмин, Л.Ю. Исмаилова и др. – Красноярск: Научно-инновационный центр, 2011. – 156 с. – ISBN 978-5-904771-17-1. – eLIBRARY ID: 19456475.

[12] Зорин П.А., Стукач О.В. Статистическое моделирование тепловых характеристик жилых домов на основе данных теплосчетчиков / Новые информационные технологии в исследовании сложных структур: материалы Тринадцатой Международной конференции. Томский государственный университет. – Томск, 07–09 сентября 2020. – С. 11. – eLIBRARY ID: 44189681.

[13] Платформа «Умное ЖКХ». [Электронный ресурс]. – 2025. – Режим доступа: https://digitaldeveloper.ru/proptech_review/tpost/0dxxm72jl1-umnoe-zhkh (дата обращения: 05.12.2025).

[14] РосКвартал. Цифровые сервисы для взаимодействия с жителями. [Электронный ресурс]. – 2025. – Режим доступа: https://roskvartal.ru (дата обращения: 05.12.2025).

[15] Автоматизация аварийно-диспетчерской службы, АДС ЖКХ. [Электронный ресурс]. – Режим доступа: https://set-r.ru/product/1s-predpriyatie/nashi-razrabotki/dispetcherskaya/ (дата обращения: 19.12.2025).

[16] Автоматизация ЖКХ. [Электронный ресурс]. – Режим доступа:http://eokr.ru/asedds/ (дата обращения: 19.12.2025).

[17] Digital transformation in housing management / McKinsey & Company [Электронный ресурс]. – 2022. – Режим доступа: https://www.mckinsey.com/industries/real-estate/our-insights/digital-transformation-in-housing (дата обращения: 12.11.2025).

[18] Дубровин М. Г. Концепция проективного мониторинга и управления объектами ИТ-инфраструктуры // ИТНОУ: Информационные технологии в науке, образовании и управлении. – 2020. – №. 1 (15). – С. 44–49.

[19] ООО «Корпорация «Парус» официальный сайт. [Электронный ресурс]. – Режим доступа: https://parus.com/ (дата обращения: 11.12.2025).

[20] Zhang Y. et al. Predictive maintenance using machine learning // Journal of Smart Cities, 2021.

[21] Платформа «Умный город». [Электронный ресурс]. – Режим доступа: https://www.rusatom-utilities.ru/activities/tsifrovye-tekhnologii/

[22] https://domopult.ru/ (дата обращения: 18.12.2025). dlya-goroda-i-regiona/smart/smart-city-base-detail/ (дата обращения: 01.12.2025).
[23] Transmission Systems for Communications, 3rd ed., Western Electric Co., Winston-Salem, NC, 1985, pp. 44–60.
[24] Домопульт – платформа для ЖКХ, управляющих компаний и ТСЖ. [Электронный ресурс]. – Режим доступа
[25] ОнлайнЖКХ: оплатить коммунальные услуги, узнать задолженность. [Электронный ресурс]. – Режим доступа: https://xn--80amifbldd4e.xn--p1ai/ (дата обращения: 19.12.2025).
[26] Профстандарт 16.018 | Специалист по управлению многоквартирным домом. [Электронный ресурс]. – Режим доступа: https://classinform.ru/profstandarty/16.018-spetcialist-po-upravleniiu-mnogokvartirnym-domom.html (дата обращения: 19.11.2025).
[27] Эффективность автоматизации ЖКХ на платформе СТЕК. [Электронный ресурс]. – Режим доступа: https://stack-it.ru/effektivnost-avtomatizaczii-na-platforme-stek/ (дата обращения: 19.12.2025).
[28] Цифровая трансформация в управлении жильем. [Электронный ресурс]. – Режим доступа: https://www.mckinsey.com/industries/real-estate/our-insights/digital-transformation-in-housing (дата обращения: 05.12.2025).
[29] Зорин П.А., Купреков С.В., Пуговкин А.В. Контроль энергоэффективности теплоснабжения зданий типовой застройки // Электронные средства и системы управления / Томский государственный университет систем управления и радиоэлектроники (Томск). – 2018. – N 1–2. – С. 302–305.
[30] Стукач О.В., Ершов И.А., Кутузов Д.В. LSTM-модель потребления тепловой энергии в многоэтажном жилом здании // Системная инженерия и инфокоммуникации. - 2025. – N 4. – С. 11–14. – https://sys-engine.ru/index.php/SEI/article/view/35.
[31] Жмудь В., Ляпидевский А., Аврамчук В., Стукач О, Рот Х. Технология промышленного интернета вещей: возможные барьеры и пути их преодоления // Автоматика и программная инженерия. – N 2(28). – 2019. – С. 50–61. – ISSN 2312-4997. – eLIBRARY ID: 39241406.
[32] Интеллектуальный анализ данных при проактивном управлении. [Электронный ресурс]. – Режим доступа: http://simulation.su/uploads/files/default/2019-voronin-evstigneev-drojjin-borovsky.pdf (дата обращения: 12.11.2025).
[33] Оптимизация энергопотребления зданий с использованием искусственного интеллекта. [Электронный ресурс]. – Режим доступа: https://www.intelvision.ru/blog/optimizacziya-energopotrebleniya-zdanij-s-ispolzovaniem-iskusstvennogo-inte (дата обращения: 20.12.2025).

Информация об авторе

Таргачева Фатима Субханвердиевна, студент кафедры АСОИУ, направления «Информатика вычислительная техника» Астраханского государственного технического университета, г. Астрахань, Россия, fatimatargacheva@gmail.com

Мамлеева Аделя Рифкатовна, старший преподаватель кафедры АСОИУ, Астраханского государственного технического университета, г. Астрахань, Россия, amamleeva@ilabsltd.com

Proactive management of residential building infrastructure: A review of a modern digital platform

Fatima Targacheva, Adelya Mamleeva

Astrakhan State Technical University, Astrakhan, Russia

Abstract **– The article examines the current state of management of residential buildings in the issues of digitalization of housing and communal services. It is noted that the dispatching services of management organizations remain a key element of operational management. However, in most cases they continue to operate within the framework of a reactive model focused on eliminating the consequences of incidents that have already occurred. The main problems of housing stock operation are analyzed, including high unplanned costs, an increase in the number of requests from residents, downtime of engineering systems, and legal risks. An overview of common dispatch platforms used in the housing and communal service has been performed, highlighting functionality and limitations. Special attention is paid to the criteria for the effectiveness of dispatch systems and the role of analytical tools. The necessity of forming a modern dispatcher workplace focused on proactive management, data integration, and intelligent decision support has been substantiated.**

Keywords **– dispatch systems, apartment building management, proactive management, housing and utilities, accident management in housing and utilities, digital housing and utilities platforms, data analytics.**

REFERENCES

[1] Rosstat. Zhilishchnyj fond Rossijskoj Federacii: statisticheskij sbornik. – M.: Rosstat, 2023. – 150 s.
[2] Federal'nyj zakon ot 21.07.2014 № 209-FZ "O gosudarstvennoj informacionnoj sisteme zhilishchno-kommunal'nogo hozyajstva" [Elektronnyj resurs]. – Rezhim dostupa: http://www.consultant.ru/document/cons_doc_LAW_165820 (data obrashcheniya: 12.11.2025).
[3] Federal'nyj zakon ot 13.12.2024 № 493-FZ \"O vnesenii izmenenij v Zhilishchnyj kodeks Rossijskoj Federacii\". – [Elektronnyj resurs]. – Rezhim dostupa: https://www.consultant.ru/document/cons_doc_LAW_493074 (data obrashcheniya: 12.11.2025).
[4] Koncepciya cifrovizacii gorodskogo hozyajstva do 2030 goda (utv. Pravitel'stvom RF) [Elektronnyj resurs]. – Rezhim dostupa: http://static.government.ru/media/files/7gWr5l5t6bTctkfqRt1f64CXhWlZv6dA.pdf (data obrashcheniya: 12.11.2025).
[5] Strategiya razvitiya stroitel'noj otrasli i zhilishchno-kommunal'nogo hozyajstva Rossijskoj Federacii na period do 2030 goda [Elektronnyj resurs]. – Rezhim dostupa: http://static.government.ru/media/files/8gWr5l5t6bTctkfqRt1f64CXhWlZv6dA.pdf (data obrashcheniya: 12.11.2025).
[6] Prokurory vyyavili bolee 424500 narushenij zakonov v sfere ZhKH v 2024 godu. – URL: https://gkhrazvitie.ru/tape/news/2025/03/prokurory-vyyavili-bolee-424-500-narushenii-zakonov-v-sfere-zhkh-v-2024-godu/ (data obrashcheniya: 19.12.2025).
[7] Dispetcher 24: avtomatizaciya dispetcherskoj sluzhby dlya ZhKH [Elektronnyj resurs]. – Rezhim dostupa: https://ds24.ru/ (data obrashcheniya: 12.11.2025).
[8] Sistemy avtomatizacii 1S ZhKH, TSZh, ZhSK [Elektronnyj resurs]. – Rezhim dostupa: https://www.1cbit.ru/1s-otrasli/avtomatizaciya-zhkkh-tszh/ (data obrashcheniya: 12.11.2025).
[9] Avtomatizirovannaya sistema upravleniya Zhilishchnyj standart [Elektronnyj resurs]. – Rezhim dostupa: https://catalog.arppsoft.ru/product/6044568 (data obrashcheniya: 12.11.2025).
[10] Proaktivnoe obsluzhivanie v sfere ZhKH kak trend tekhnologicheskogo razvitiya. [Elektronnyj resurs]. – Rezhim dostupa: https://cyberleninka.ru/article/n/proaktivnoe-obsluzhivanie-v-sfere-zhkh-kak-trend-tehnologicheskogo-razvitiya (data obrashcheniya: 09.12.2025).

[11] Informacionnye sistemy i tekhnologii: monografiya / O.I. Babina, N.Yu. Dyumin, L.Yu. Ismailova i dr. – Krasnoyarsk: Nauchno-innovacionnyj centr, 2011. – 156 s. – ISBN 978-5-904771-17-1. – eLIBRARY ID: 19456475.
[12] Zorin P.A., Stukach O.V. Statisticheskoe modelirovanie teplovyh harakteristik zhilyh domov na osnove dannyh teploschetchikov / Novye informacionnye tekhnologii v issledovanii slozhnyh struktur: materialy Trinadcatoj Mezhdunarodnoj konferencii. Tomskij gosudarstvennyj universitet. – Tomsk, 07–09 sentyabrya 2020. – S. 11. – eLIBRARY ID: 44189681.
[13] Platforma «Umnoe ZhKH». Elektronnyj resurs]. – 2025. – Rezhim dostupa: https://digitaldeveloper.ru/proptech_review/tpost/0dxxm72jl1-umnoe-zhkh (data obrashcheniya: 05.12.2025).
[14] RosKvartal. Cifrovye servisy dlya vzaimodejstviya s zhitelyami. [Elektronnyj resurs]. – 2025. – Rezhim dostupa: https://roskvartal.ru (data obrashcheniya: 05.12.2025).
[15] Avtomatizaciya avarijno-dispetcherskoj sluzhby, ADS ZhKH. [Elektronnyj resurs]. – Rezhim dostupa: https://set-r.ru/product/1s-predpriyatie/nashi-razrabotki/dispetcherskaya/ (data obrashcheniya: 19.12.2025).
[16] Avtomatizaciya ZhKH. [Elektronnyj resurs]. – Rezhim dostupa:http://eokr.ru/asedds/ (data obrashcheniya: 19.12.2025).
[17] Digital transformation in housing management / McKinsey & Company [Электронный ресурс]. – 2022. – Режим доступа: https://www.mckinsey.com/industries/real-estate/our-insights/digital-transformation-in-housing (дата обращения: 12.11.2025).
[18] Dubrovin M. G. Koncepciya proektivnogo monitoringa i upravleniya ob"ektami IT-infrastruktury // ITNOU: Informacionnye tekhnologii v nauke, obrazovanii i upravlenii. – 2020. – №. 1 (15). – S. 44–49.
[19] OOO «Korporaciya «Parus» oficial'nyj sajt. [Elektronnyj resurs]. – Rezhim dostupa: https://parus.com/ (data obrashcheniya: 11.12.2025).
[20] Zhang Y. et al. Predictive maintenance using machine learning // Journal of Smart Cities, 2021.
[21] Platforma «Umnyj gorod». [Elektronnyj resurs]. – Rezhim dostupa: https://www.rusatom-utilities.ru/activities/tsifrovye-tekhnologii/
[22] https://domopult.ru/ (дата обращения: 18.12.2025). dlya-goroda-i-regiona/smart/smart-city-base-detail/ (дата обращения: 01.12.2025).
[23] Transmission Systems for Communications, 3rd ed., Western Electric Co., Winston-Salem, NC, 1985, pp. 44–60.
[24] Domopul't – platforma dlya ZhKH, upravlyayushchih kompanij i TSZh. [Elektronnyj resurs]. – Rezhim dostupa
[25] OnlajnZhKH: oplatit' kommunal'nye uslugi, uznat' zadolzhennost'. [Elektronnyj resurs]. – Rezhim dostupa: https://xn--80amifbldd4e.xn--p1ai/ (data obrashcheniya: 19.12.2025).
[26] Profstandart 16.018 | Specialist po upravleniyu mnogokvartirnym domom. [Elektronnyj resurs]. – Rezhim dostupa: https://classinform.ru/profstandarty/16.018-spetcialist-po-upravleniiu-mnogokvartirnym-domom.html (data obrashcheniya: 19.11.2025).
[27] Effektivnost' avtomatizacii ZhKH na platforme STEK. [Elektronnyj resurs]. – Rezhim dostupa: https://stack-it.ru/effektivnost-avtomatizaczii-na-platforme-stek/ (data obrashcheniya: 19.12.2025).
[28] Cifrovaya transformaciya v upravlenii zhil'em. [Elektronnyj resurs]. – Rezhim dostupa: https://www.mckinsey.com/industries/real-estate/our-insights/digital-transformation-in-housing (data obrashcheniya: 05.12.2025).
[29] Zorin P.A., Kuprekov S.V., Pugovkin A.V. Kontrol' energoeffektivnosti teplosnabzheniya zdanij tipovoj zastrojki // Elektronnye sredstva i sistemy upravleniya / Tomskij gosudarstvennyj universitet sistem upravleniya i radioelektroniki (Tomsk). – 2018. – N 1–2. – S. 302–305.
[30] Stukach O.V., Ershov I.A., Kutuzov D.V. LSTM-model' potrebleniya teplovoj energii v mnogoetazhnom zhilom zdanii // Sistemnaya inzheneriya i infokommunikacii. - 2025. – N 4. – S. 11–14. – https://sys-engine.ru/index.php/SEI/article/view/35.
[31] Zhmud' V., Lyapidevskij A., Avramchuk V., Stukach O, Rot H. Tekhnologiya promyshlennogo interneta veshchej: vozmozhnye bar'ery i puti ih preodoleniya // Avtomatika i programmnaya inzheneriya. – N 2(28). – 2019. – S. 50–61. – ISSN 2312-4997. – eLIBRARY ID: 39241406.
[32] Intellektual'nyj analiz dannyh pri proaktivnom upravlenii. [Elektronnyj resurs]. – Rezhim dostupa: http://simulation.su/uploads/files/default/2019-voronin-evstigneev-drojjin-borovsky.pdf (data obrashcheniya: 12.11.2025).
[33] Optimizaciya energopotrebleniya zdanij s ispol'zovaniem iskusstvennogo intellekta. [Elektronnyj resurs]. – Rezhim dostupa: https://www.intelvision.ru/blog/optimizacziya-energopotrebleniya-zdanij-s-ispolzovaniem-iskusstvennogo-inte (data obrashcheniya: 20.12.2025).

Information about the authors

Targacheva Fatima Subhanverdievna, student of the AIPCS Department, Computer Science and Engineering field of study, Astrakhan State Technical University, Astrakhan, Russia, fatimatargacheva@gmail.com

Mamleeva Adelya Rifkatovna, senior lecturer of the AIPCS Department, Astrakhan State Technical University, Astrakhan, Russia, amamleeva@ilabsltd.com

Системная инженерия и инфокоммуникации

Научный сетевой журнал «Системная инженерия и инфокоммуникации» (свид. о регистрации СМИ Эл № ФС77-88783 от 22 ноября 2024 г., ISSN: 3034-686X) — это некоммерческое рецензируемое сетевое издание со свободным доступом. Сайт журнала: https://sys-engine.ru

Журнал ориентирован в первую очередь на молодых специалистов в области информационных технологий, робототехники, автоматики, электроники и телекоммуникаций. Аспиранты, магистранты и студенты могут опубликовать в журнале результаты своих исследований, собственный опыт разработки проектов, обзорные статьи.

Журнал выходит четыре раза в год, и принимает к публикации оригинальные статьи, соответствующие тематике журнала. Публикация в журнале бесплатна.

Журнал не преследует коммерческой выгоды, не выплачивает вознаграждения авторам, редакция и рецензенты не получают оплату за свою работу. Публикуя статью в сетевом издании «Системная инженерия и инфокоммуникации», авторы присоединяются к цели журнала — свободному распространению научного знания.

Свидетельство о регистрации СМИ Эл № ФС77-88783 от 22 ноября 2024 г.
ISSN: 3034-686X
Учредитель: Кутузов Денис Валерьевич
Сетевое издание размещено по адресу: sys-engine.ru
Электронная почта: d_kutuzov@mail.ru
Телефон: +7(964)88-22-445