НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
сентябрь–октябрь 2026 Том 26 № 5 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
September–October 2026 Vol. 26 No 5 http://ntv.ifmo.ru/en/
ISSN 2226-1494 (print) ISSN 2500-0373 (online)

НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК
ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ

doi: 10.17586/2226-1494-2026-26-5-1009-1017
УДК 004.056

Методика интегральной оценки качества моделей машинного обучения для обнаружения кибератак в сетевом трафике

Алексей Александрович Лукоянов[1]✉, Евгений Алексеевич Рогозин[2], Руслан Евгеньевич Рогозин[3], Илья Иосифович Лившиц[4]

[1] Российский государственный гуманитарный университет, Москва, 125047, Российская Федерация
[1] АНО «Проектный офис по развитию туризма и гостеприимства Москвы», Москва, 125009, Российская Федерация
[2] Воронежский институт МВД России, Воронеж, 394065, Российская Федерация
[3] МИРЭА — Российский технологический университет, Москва, 119454, Российская Федерация
[4] Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация

[1] 9261781233@mail.ru✉, https://orcid.org/0009-0001-4519-6882
[2] evgenirogozin@yandex.ru, https://orcid.org/0000-0002-4455-7535
[3] ruslan-96-01-09@mail.ru, https://orcid.org/0000-0002-2913-6038
[4] Livshitz.il@yandex.ru, https://orcid.org/0000-0003-0651-8591

Аннотация

Введение. Модели машинного обучения активно применяются для обнаружения кибератак на основе анализа сетевого трафика. Традиционные метрики оценки качества не учитывают влияние временных характеристик данных и могут скрывать снижение качества в отдельных временных интервалах, а также при изменении интенсивности трафика. Подход Key Performance Indicators, взвешенное геометрическое агрегирование и байесовские методы позволяют объединять отдельные характеристики качества, однако сами по себе не учитывают его изменение между заранее заданными состояниями функционирования модели. **Метод.** Разработанная методика интегральной оценки учитывает частоту и уверенность ошибочных решений, а также снижение качества модели в конкретных состояниях функционирования. Частный показатель качества модели рассчитывается для каждой комбинации временного интервала, целевого уровня полноты и режима интенсивности трафика. Показатель формируется путем мультипликативного объединения классификационной оценки, средней калиброванной уверенности в правильных решениях, потери от ошибочных решений и временного множителя. Интегральный показатель получают усреднением частных показателей пропорционально числу наблюдений и последующей коррекцией по величине отклонений ниже средневзвешенного уровня. Экспериментальные выборки сформированы с использованием единого хронологического разделения данных, исключающего использование тестовых записей при обучении моделей, калибровке вероятностей и выборе порогов классификации. **Основные результаты.** Методика апробирована на наборе данных GothamDataset2025. Тестовая выборка включала 1 777 237 записей сетевого трафика. Проведено сравнение моделей логистической регрессии, случайного леса и гистограммного градиентного бустинга в 18 состояниях. При порогах, обеспечивающих на валидационной выборке полноту обнаружения атак не ниже 0,50, случайный лес показал наибольшие значения сбалансированной точности — 0,693933 и коэффициента корреляции Мэтьюса — 0,466685. Наибольшая интегральная оценка качества, равная 0,264163, получена для гистограммного градиентного бустинга. Предложенная методика выявила состояния со сниженным и нулевым значениями частного показателя качества, которые не обнаруживались при анализе сбалансированной точности и коэффициента корреляции Мэтьюса, рассчитанных по всей тестовой выборке, и изменила итоговое ранжирование моделей. **Обсуждение.** Для применения методики до начала оценки необходимо определить набор рассматриваемых состояний и значения параметров агрегирования. Методика может применяться для выбора моделей обнаружения кибератак при изменяющихся условиях сетевого трафика.

Ключевые слова

интегральная оценка качества, машинное обучение, обнаружение кибератак, сетевой трафик, информационная безопасность, калибровка вероятностей, частный показатель качества

© Лукоянов А.А., Рогозин Е.А., Рогозин Р.Е., Лившиц И.И., 2026

Ссылка для цитирования: Лукоянов А.А., Рогозин Е.А., Рогозин Р.Е., Лившиц И.И. Методика интегральной оценки качества моделей машинного обучения для обнаружения кибератак в сетевом трафике // Научно-технический вестник информационных технологий, механики и оптики. 2026. Т. 26, № 5. С. 1009–1017. doi: 10.17586/2226-1494-2026-26-5-1009-1017

Integral quality assessment of machine learning models for cyberattack detection in network traffic

Aleksey A. Lukoyanov[1]✉, Evgeny A. Rogozin[2], Ruslan E. Rogozin[3], Ilya I. Livshits[4]

[1] Russian State University for the Humanities (RSUH), Moscow, 125047, Russian Federation
[1] Autonomous Non-Profit Organization "Project Office for Tourism and Hospitality Development of Moscow", Moscow, 125009, Russian Federation
[2] Voronezh Institute of the Ministry of Internal Affairs of Russia, Voronezh, 394065, Russian Federation
[3] MIREA — Russian Technological University, Moscow, 119454, Russian Federation
[4] ITMO University, Saint Petersburg, 197101, Russian Federation

[1] 9261781233@mail.ru✉, https://orcid.org/0009-0001-4519-6882
[2] evgenirogozin@yandex.ru, https://orcid.org/0000-0002-4455-7535
[3] ruslan-96-01-09@mail.ru, https://orcid.org/0000-0002-2913-6038
[4] Livshitz.il@yandex.ru, https://orcid.org/0000-0003-0651-8591

Abstract

Machine learning models are widely used to detect cyberattacks through the analysis of network traffic. Traditional performance metrics do not account for the temporal characteristics of the data and may mask performance degradation within individual time intervals or under varying traffic intensity. Approaches based on key performance indicators, weighted geometric aggregation, and Bayesian methods allow individual performance characteristics to be combined; however, by themselves, they do not account for changes in model performance across predefined operating states. The developed integrated assessment methodology takes into account the frequency and certainty of erroneous decisions, as well as the degradation of model quality in specific operating states. A partial model quality indicator is calculated for each combination of time interval, target recall level, and traffic intensity mode. The indicator is formed by multiplicatively combining the classification score, the average calibrated confidence in correct decisions, the loss from erroneous decisions, and a time multiplier. The integrated indicator is obtained by averaging the partial indicators proportionally to the number of observations and subsequently adjusting for deviations below the weighted average. Experimental samples were generated using a single chronological data partition, eliminating the use of test records during model training, probabilistic calibration, and selection of classification thresholds. The methodology was evaluated on GothamDataset2025 using a test set of 1,777,237 network traffic records. Logistic regression, random forest, and histogram-based gradient boosting were compared across 18 operating states. At thresholds selected to provide validation recall of at least 0.50, random forest achieved the highest balanced accuracy of 0.693933 and the Matthews correlation coefficient of 0.466685. Histogram-based gradient boosting obtained the highest integral quality score of 0.264163. The methodology identified states with reduced or zero state-specific indicators that were not revealed by balanced accuracy and the Matthews correlation coefficient calculated over the entire test set, and changed the final model ranking. Operating states and aggregation parameters must be defined before assessment. The methodology supports the selection of cyberattack detection models under varying network traffic conditions.

Keywords

integral quality assessment, machine learning, cyberattack detection, network traffic, information security, probability calibration, state-specific quality indicator

For citation: Lukoyanov A.A., Rogozin E.A., Rogozin R.E., Livshits I.I. Integral quality assessment of machine learning models for cyberattack detection in network traffic. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2026, vol. 26, no. 5, pp. 1009–1017 (in Russian). doi: 10.17586/2226-1494-2026-26-5-1009-1017

Введение

Модели машинного обучения применяются для обнаружения кибератак в сетевом трафике [1]. Качество таких моделей обычно оценивают с помощью положительной прогностической ценности, полноты обнаружения, специфичности, сбалансированной точности и коэффициента корреляции Мэтьюса [2]. Однако показатели, вычисленные по совокупности тестовых данных, могут скрывать отсутствие срабатываний модели в отдельных временных интервалах, а также снижение качества при изменении интенсивности сетевого трафика. Дополнительными источниками искажения результатов являются случайное распределение зависимых записей между обучающей и тестовой выборками [3] и изменение распределения данных во времени [4, 5].

Работы [6–8] посвящены точности и объяснимости моделей искусственного интеллекта в системах управления событиями информационной безопасности, устойчивости систем обнаружения вторжений к целенаправленным воздействиям на модели машинного обучения и ресурсной эффективности обнаружения сетевых атак. Вместе с тем в них не решается задача объединения результатов, полученных для разных временных интервалов, порогов классификации и режимов интенсивности трафика. В работах [9, 10] сформулированы требования к безопасному применению систем искусственного интеллекта и управлению ими, но не

предложен количественный показатель для сопоставления моделей обнаружения кибератак при изменении условий их функционирования.

Для формирования единой оценки может использоваться подход Key Performance Indicators (KPI), основанный на взвешенной сумме нормированных показателей. Его преимуществами являются простота расчета и возможность учитывать различную значимость показателей. Вместе с тем аддитивная свертка допускает компенсацию низкого значения одного показателя высокими значениями других. Взвешенное геометрическое среднее ограничивает такую компенсацию [11], однако при расчете показателей по матрице ошибок, сформированной для всей выборки, также не отражает различия качества модели между отдельными условиями функционирования.

Байесовский подход позволяет оценить условную вероятность наличия кибератаки при отнесении анализируемого трафика моделью к легитимному классу. Такой показатель имеет вероятностную интерпретацию, однако зависит от априорной доли атак и характеризует только один вид ошибочного решения. Кроме того, он не объединяет различия результатов между временными интервалами, время обработки трафика и качество вероятностных оценок модели, для корректного использования которых требуется калибровка [12].

Таким образом, KPI- и байесовский подходы характеризуют модель по показателям, рассчитанным для выборки в целом, и не позволяют определить, в каких сочетаниях временного интервала, целевого уровня полноты и интенсивности трафика возникает снижение качества. Для решения этой задачи необходимо оценивать модель отдельно в каждом заданном состоянии и учитывать неблагоприятные результаты при формировании итогового показателя.

В настоящей работе представлена новая методика интегральной оценки качества моделей машинного обучения для обнаружения кибератак при изменении временного интервала наблюдения, целевого уровня полноты и режима интенсивности сетевого трафика.

Научная новизна методики состоит в способе формирования интегральной оценки качества модели. Для каждого сочетания временного интервала, целевого уровня полноты и режима интенсивности трафика рассчитывается частный показатель качества. Затем полученные значения объединяются с учетом числа наблюдений и корректируются по величине отклонений ниже средневзвешенного уровня.

Описание предлагаемой методики

Рассмотрим задачу бинарной классификации записей сетевого трафика на вредоносный и легитимный классы. Вредоносный трафик относится к положительному классу, легитимный — к отрицательному.

Формирование условий оценки. Пусть Q — множество сравниваемых моделей, а $q \in Q$ — индекс модели. Модели сравниваются при одинаковых целевых уровнях полноты, поскольку числовые значения порогов зависят от шкалы выходных оценок каждой модели. Пороги определяются на валидационной выборке. Множество оцениваемых состояний Ω задается декартовым произведением:

$$\Omega = T \times R \times \Psi,\ \omega = (t, r, \psi),$$

где множества: T — временных блоков наблюдения, R — целевых уровней полноты, Ψ — режимов интенсивности сетевого трафика; $\omega = (t, r, \psi) \in \Omega$ — отдельное состояние, задаваемое соответствующим сочетанием факторов.

Для модели q и целевого уровня полноты r порог классификации модели q определяется:

$$\theta_{q,r} = \max\{\theta: TPR_q^{\text{val}}(\theta) \geq r\},$$

где θ — проверяемое значение порога; TPR_q^{val} — полнота обнаружения атак моделью q, рассчитанная на валидационной выборке при пороге θ; val — валидационная выборка.

Таким образом, для каждой модели устанавливается максимальное значение порога, при котором достигается заданный уровень полноты обнаружения атак.

Классификационная составляющая. В каждом состоянии формируется матрица ошибок, по которой рассчитываются показатели качества:

$$PPV_\omega = \frac{TP_\omega}{TP_\omega + FP_\omega},\quad TPR_\omega = \frac{TP_\omega}{TP_\omega + FN_\omega},\quad TNR_\omega = \frac{TN_\omega}{TN_\omega + FP_\omega}, \tag{1}$$

где TP_ω и FN_ω — число правильно обнаруженных и пропущенных вредоносных записей; TN_ω и FP_ω — число легитимных записей, правильно распознанных и ошибочно отнесенных к вредоносному классу. Показатели PPV_ω, TPR_ω, TNR_ω обозначают соответственно положительную прогностическую ценность, полноту обнаружения атак и специфичность; FNR_ω, FPR_ω — доли пропусков атак и ложных срабатываний.

Частный показатель качества модели рассчитывается только для состояний, содержащих записи обоих фактических классов. При отсутствии одного из классов состояние исключается из агрегирования, а веса остальных состояний нормируются повторно. Если в состоянии присутствуют фактически вредоносные записи, но модель не отнесла ни одной записи к вредоносному классу, $TPR_q(\omega) = 0$, а неопределенное значение $PPV_q(\omega)$ принимается равным нулю. В результате частный показатель качества модели $k_q(\omega) = 0$. Мультипликативная классификационная оценка качества модели q в состоянии ω имеет вид:

$$Q_q(\omega) = PPV_\omega^\alpha \times TPR_\omega^\beta \times TNR_\omega^\gamma, \tag{2}$$

где α, β, γ — показатели степени, определяющие значимость PPV_ω, TPR_ω, TNR_ω соответственно. В эксперименте принято $\alpha = \beta = \gamma = 1$, поэтому три характеристики оказывают одинаковое влияние. Нулевое значение любой из них приводит к $Q_q(\omega) = 0$ и не компенсируется высокими значениями остальных характеристик.

Учет уверенности модели в правильных и ошибочных решениях. Пусть $\hat{p}_i \in [0, 1]$ — откалибро-

ванная вероятность принадлежности объекта i к вредоносному классу, а $\hat{y}_i \in \{0, 1\}$ — решение модели. Уверенность модели в принятом решении для записи i определяется по формуле:

$$d_i = \begin{cases} \hat{p}_i, \hat{y}_i = 1 \\ 1 - \hat{p}_i, \hat{y}_i = 0 \end{cases}, \quad (3)$$

где $\hat{y}_i = 1$ и $\hat{y}_i = 0$ — значение вредоносного и легитимного классов.

Для правильно классифицированных записей состояния ω определяется среднее значение калиброванной уверенности модели:

$$\bar{d}_{\omega,\text{cor}} = \frac{1}{n_{\omega,\text{cor}}} \sum_{i \in I_{\omega,\text{cor}}} d_i, \quad (4)$$

где $I_{\omega,\text{cor}}$ — множество правильно классифицированных записей состояния ω; $n_{\omega,\text{cor}} = |I_{\omega,\text{cor}}|$ — количество таких записей. При отсутствии правильных решений значение $\bar{d}_{\omega,\text{cor}}$ принимается равным нулю. Аналогично определяются средние значения калиброванной уверенности для ложноотрицательных и ложноположительных решений.

Ошибочные решения учитываются посредством частотно-взвешенной потери:

$$L_{\omega,\text{err}} = \eta FNR_\omega \bar{d}_{FN,\omega} + (1 - \eta) FPR_\omega \bar{d}_{FP,\omega}, \quad (5)$$

где $FNR_\omega = FN_\omega / TP_\omega + FN_\omega$, $FPR_\omega = FP_\omega / TN_\omega + FP_\omega$; $\bar{d}_{FN,\omega}$ и $\bar{d}_{FP,\omega}$ — средние значения калиброванной уверенности модели соответственно в ложноотрицательных и ложноположительных решениях (при отсутствии ошибок соответствующее среднее принимается равным нулю); $\eta \in [0, 1]$ — относительная значимость пропусков атак.

В каждом слагаемом выражения (5) первый множитель характеризует частоту ошибки, а второй — уверенность модели в ошибочном решении. По этой причине значение потери увеличивается как при росте частоты ошибок, так и при повышении уверенности модели в них. Нормирование по числу объектов соответствующего фактического класса уменьшает зависимость результата от различий в размерах классов.

Оценка временной эффективности. Время обработки преобразуется в безразмерный множитель:

$$\tilde{t}_q = \frac{\bar{t}_q}{T_{\text{ref}}}, \; V_q = \exp[-\lambda \min(\tilde{t}_q, \tau_{\max})], \; \lambda = -\ln q_t, \quad (6)$$

где $\tilde{t}_q$ — безразмерное нормированное время обработки; $\bar{t}_q$ — медианное сквозное время обработки одной записи моделью q; T_{ref} — опорное время обработки; $q_t \in (0, 1)$ — значение временного множителя V_q при $\tilde{t}_q = 1$; $\tau_{\max}$ — безразмерная верхняя граница нормированного времени, ограничивающая величину временного штрафа. С увеличением времени обработки значение V_q уменьшается.

Множитель V_q является безразмерной величиной, поскольку определяется через безразмерное нормированное время $\tilde{t}_q$.

Частный показатель качества. Объединение классификационной, вероятностной и временной составляющих образует частный показатель качества модели q в состоянии ω:

$$k_q(\omega) = Q_q(\omega) V_q \left(\frac{1 + \bar{d}_{\omega,\text{cor}}}{2} \right)^a (1 - L_{\omega,\text{err}})^b, \quad (7)$$

где $Q_q(\omega)$ — классификационная составляющая; $a \geq 0$, $b \geq 0$ — показатели степени, регулирующие влияние уверенности правильных решений и потерь от ошибок соответственно.

Согласно формуле (7) $k_q(\omega) \in [0; 1]$, причем большее значение соответствует более высокому качеству модели в состоянии ω. Индекс равен нулю при отказе одной из обязательных классификационных характеристик или максимальной потере от ошибочных решений. Преобразование $(1 + \bar{d}_{\omega,\text{cor}})/2$ ограничивает влияние уверенности правильных решений диапазоном [0,5; 1].

Формирование итоговой оценки. Для каждой модели рассчитывается частный показатель качества $k_q(\omega)$ во всех состояниях $\omega \in \Omega$. В эксперименте использованы три временных блока, три целевых уровня полноты и два режима интенсивности трафика, образующие 18 состояний. Такое разбиение позволяет оценить изменение качества во времени и при разных рабочих точках модели, а также сопоставить результаты для двух уровней нагрузки. Количество уровней каждого фактора выбирается с учетом цели исследования и объема данных, достаточного для расчета показателей качества в каждом рассматриваемом состоянии.

Для сравнения моделей частные показатели качества $k_q(\omega)$, рассчитанные для отдельных состояний, объединяются в средневзвешенную оценку. Обозначим через $\Omega^* = \{\omega \in \Omega: TP_\omega + FN_\omega > 0, TN_\omega + FP_\omega > 0\}$ множество состояний, содержащих записи обоих фактических классов. Поскольку Ω^* формируется отбором состояний из исходного множества Ω, выполняется $\Omega^* \subseteq \Omega$. Частные показатели качества и итоговая интегральная оценка рассчитываются по состояниям $\omega \in \Omega^*$. При этом Ω^* — только те состояния из Ω, в которых присутствуют записи обоих фактических классов и можно корректно рассчитать TPR и TNR.

Общая средневзвешенная оценка $\bar{K}_q$ и нормированный вес w_ω состояния ω имеют вид:

$$\bar{K}_q = \sum_{\omega \in \Omega^*} w_\omega k_q(\omega), \quad (8)$$

$$w_\omega = \frac{n_\omega}{\sum_{\nu \in \Omega^*} n_\nu}, \quad \sum_{\omega \in \Omega^*} w_\omega = 1,$$

где n_ω — число наблюдений в состоянии ω; ν — вспомогательный индекс суммирования по состояниям множества Ω^*.

Согласно формуле (8), состояния с большим числом наблюдений сильнее влияют на итоговый результат. Одинаковое значение $\bar{K}_q$ может быть получено как при близких частных результатах, так и при наличии отдельных состояний с низким качеством. Рассчитаем среднеквадратичное отклонение (Δ_q) только тех частных показателей качества, которые ниже средневзвешенного уровня. Учет только неблагоприятных отклонений соответствует подходу к анализу риска, основанному на отклонениях ниже целевого значения [13].

$$\Delta_q^- = \sqrt{\sum_{\omega \in \Omega^*} w_\omega [\max(0, \overline{K}_q - k_q(\omega))]^2}. \quad (9)$$

Итоговая интегральная оценка определяется выражением:

$$K_q^{\text{int}} = \frac{\overline{K}_q}{1 + c\Delta_q^-}, \quad (10)$$

где $c \geq 0$ — коэффициент чувствительности итоговой оценки к отклонениям частных показателей качества ниже средневзвешенного уровня. Значение c задается до итогового тестирования: при $c = 0$ штраф отсутствует, а увеличение c усиливает снижение итоговой оценки. Если $\Delta_q^- = 0$, значение c не влияет на результат; при $c \to \infty$ и $\Delta_q^- > 0$ значение K_q^{int} стремится к нулю.

Штраф уменьшает итоговую оценку модели при наличии состояний, частные показатели качества которых ниже средневзвешенного уровня; величина уменьшения определяется весом и глубиной этих отклонений. Показатель K_q^{int} предназначен для сравнительного ранжирования моделей и не имеет вероятностной интерпретации.

Порядок расчета. Для применения методики выполняются следующие действия.

1. На валидационной выборке калибруются вероятностные оценки модели и определяются пороги классификации, соответствующие заданным уровням полноты.
2. До начала оценки задается множество состояний функционирования модели. Каждое состояние определяется сочетанием временного интервала наблюдения, целевого уровня полноты и режима интенсивности сетевого трафика. Границы интервалов и режимов устанавливаются единообразно для всех сравниваемых моделей.
3. Для каждого состояния формируется матрица ошибок и определяется число относящихся к нему наблюдений.
4. По формулам (1)–(6) рассчитываются классификационные, вероятностные и временные характеристики состояния.
5. По (7) определяется частный показатель качества модели в каждом состоянии.
6. По (8)–(10) частные показатели качества объединяются в средневзвешенную оценку, рассчитывается неблагоприятное отклонение и формируется итоговая интегральная оценка.
7. Для каждой модели формируется итоговая интегральная оценка K_q^{int}. Модели ранжируются по убыванию ее значения, а частные показатели качества $k_q(\omega)$ анализируются для выявления условий, в которых наблюдается снижение качества.

Апробация методики

В табл. 1 приведены основные параметры методики, использованные в эксперименте.

Параметры методики устанавливаются до проведения итогового тестирования и выбираются для конкретной задачи. При наличии модели угроз значения α, β, γ и η определяются исходя из значимости показателей и возможных последствий ошибок. В проведенном эксперименте параметрам классификационного блока присвоены равные значения, а ошибки учитывались симметрично.

Экспериментальные данные и модели. Из обработанных файлов GothamDataset2025[1] [14] сформированы обучающая, валидационная и тестовая выборки объемом 320 000, 400 000 и 1 777 237 записей соответственно. В тестовой выборке 995 174 записи от-

[1] Belarbi O., Spyridopoulos T., Anthi E., Rana O., Carnelli P., Khan A. A Device-Level IoT Network Traffic Dataset with Distributed Capture and Non-IID Characteristics [Dataset]. Zenodo, 2025 [Электронный ресурс]. Режим доступа: https://zenodo.org/records/14502760 (дата обращения: 30.09.2026). doi: 10.5281/zenodo.14502760.

Таблица 1. Параметры методики и значения, использованные в эксперименте

Table 1. Methodological parameters and values used in the experiment

Параметр	Назначение	Значение в эксперименте и обоснование
α, β, γ	Показатели степени, определяющие значимость *PPV*, *TPR* и *TNR*	α = β = γ =1, обеспечивают симметричную форму мультипликативной свертки без априорного усиления одной из трех характеристик
T_{ref}, q_t, $\tau_{\max}$	Параметры учета времени обработки	T_{ref} = 1 мс; q_t = 0,90; $\tau_{\max}$ = 2. Опорное время соответствует производительности 1000 последовательно обрабатываемых записей в секунду и задается до итогового тестирования. При $\bar{t}_q = T_{\text{ref}}$ временной множитель составляет 0,90; ограничение $\tau_{\max}$ = 2 предотвращает неограниченное увеличение временного штрафа
a, *b*	Влияние уверенности в правильных решениях и потерь от ошибочных решений	*a* = *b* = 0,2. Равные малые степени выбраны для ограничения влияния вероятностных составляющих: для множителя уверенности правильных решений минимальное значение после преобразования составляет $0{,}5^{0{,}2} = 0{,}871$, поэтому он снижает частный показатель качества модели не более чем на 12,9 %
η	Соотношение пропусков атак и ложных срабатываний	η = 0,5. Пропуски атак и ложные срабатывания учитываются одинаково, поскольку данные о различии последствий этих ошибок отсутствуют
c	Влияние отклонений частных показателей качества на итоговую оценку	*c* = 1. Таким образом, величина отклонения при расчете итоговой оценки не изменяется

носились к вредоносному, а 782 063 — к легитимному трафику. Выборки сформированы в хронологическом порядке из данных, собранных 14, 15, 17 и 18 января 2025 года. Тестовые записи не использовались при обучении моделей, предварительной обработке данных, калибровке вероятностей и выборе порогов классификации. Сравнивались логистическая регрессия ($C = 1$), случайный лес (200 деревьев, максимальная глубина 20, не менее 5 записей в листе) и гистограммный градиентный бустинг (180 итераций, не более 31 листа, скорость обучения 0,08). Параметры моделей фиксировались до итогового тестирования.

На валидационной выборке выполнялась калибровка выходных вероятностей моделей. Для логистической регрессии и гистограммного градиентного бустинга применялся изотонический метод, для случайного леса — сигмоидальный. Для каждой модели выбирались пороги, обеспечивающие полноту обнаружения атак не ниже 0,25, 0,50 и 0,75.

Тестовый период 18 января 2025 года разделялся на три равных по продолжительности временных блока:

$$T_1 = [21 : 19 : 24;\ 21 : 35 : 44),$$
$$T_2 = [21 : 35 : 44;\ 21 : 52 : 04),$$
$$T_3 = [21 : 52 : 04;\ 22 : 08 : 24].$$

Временные блоки заданы непересекающимися интервалами: для T_1 и T_2 правая граница не включается, а для T_3 конечная правая граница включается.

Интенсивность определялась числом пакетных записей за одну секунду. Значения не выше медианы валидационной выборки, равной 282 пакетам/с, относились к режиму пониженной интенсивности, остальные — к режиму повышенной интенсивности. Тестовые данные распределялись по 18 ранее определенным состояниям множества Ω; для каждого состояния рассчитывались матрица ошибок и частный показатель качества модели. Все 18 состояний содержали вредоносные и легитимные записи, поэтому в эксперименте $\Omega^* = \Omega$.

Медианное время обработки одной записи составило 0,00196, 0,00260 и 0,00372 мс для логистической регрессии, случайного леса и гистограммного градиентного бустинга соответственно. При $T_{\text{ref}} = 1$ мс соответствующие значения временного множителя составили 0,999794, 0,999726 и 0,999608. Значения временного множителя для всех моделей превышали 0,9996. Таким образом, при выбранном опорном времени временная составляющая практически не повлияла на значения интегральной оценки и порядок моделей.

Расчет интегральной оценки. Для сопоставления с предлагаемой методикой по всей исследуемой выборке дополнительно были рассчитаны сбалансированная точность и коэффициент корреляции Мэтьюса [2]. Для каждой модели использовался собственный порог, обеспечивавший на валидационной выборке полноту обнаружения атак не ниже 0,50. Сбалансированная точность $BA = (TPR + TNR)/2$ характеризует среднее качество распознавания вредоносного и легитимного трафиков. Коэффициент корреляции Мэтьюса $MCC \in [-1; 1]$ учитывает все элементы матрицы ошибок; большее значение соответствует более высокому качеству классификации. В табл. 2 приведены показатели качества, рассчитанные по всей исследуемой выборке, и результаты интегральной оценки моделей.

Для проверки влияния временной составляющей выполнен дополнительный расчет при $V_q = 1$. Интегральные оценки логистической регрессии, случайного леса и гистограммного градиентного бустинга составили соответственно 0,102899, 0,100660 и 0,264257.

Модель случайный лес показала наибольшие значения сбалансированной точности $BA = 0{,}693933$ и коэффициента корреляции Мэтьюса $MCC = 0{,}466685$. Применение предложенной методики изменило ранжирование: наибольшее значение $K_q^{\text{int}} = 0{,}264163$ получено для гистограммного градиентного бустинга.

Различие связано с нулевыми значениями частного показателя у логистической регрессии и случайного леса при целевом уровне полноты 0,25. В отдельных состояниях эти модели не отнесли ни одной записи к вредоносному классу. В таких случаях $PPV_\omega(\omega)$ не определялась и в соответствии с принятым правилом принималась равной нулю, что приводило к $k_q(\omega) = 0$. Для гистограммного градиентного бустинга выполнялось $k_q(\omega) > 0$ во всех 18 состояниях.

Значения частных показателей качества $k_q(\omega)$, рассчитанные для 18 состояний, представлены на рисунке. Номера состояний соответствуют сочетаниям трех временных интервалов, трех целевых уровней полноты и двух режимов интенсивности трафика. В каждой паре состояний первый номер соответствует низкой, второй — высокой интенсивности трафика. Для каждого состояния приведена группа столбцов, соответствующих трем исследуемым моделям.

Частные показатели качества объединены в обобщенный показатель по формуле (8). Затем по формуле (9) рассчитано нижнее отклонение, а по формуле (10) получена итоговая интегральная оценка.

Обсуждение

Сопоставление с базовыми подходами. В качестве базовых подходов для сравнения использованы модели

Таблица 2. Рассчитанные показатели по исследуемой выборке и интегральные оценки моделей
Table 2. Calculated metrics for the study sample and integral model assessments

Модель	BA	MCC	$\overline{K}_q$	Δ_q^-	K_q^{int}
Логистическая регрессия	0,689779	0,460138	0,112172	0,090319	0,102880
Случайный лес	**0,693933**	**0,466685**	0,109782	0,090896	0,100634
Гистограммный градиентный бустинг	0,687158	0,456503	**0,291264**	0,102592	**0,264163**

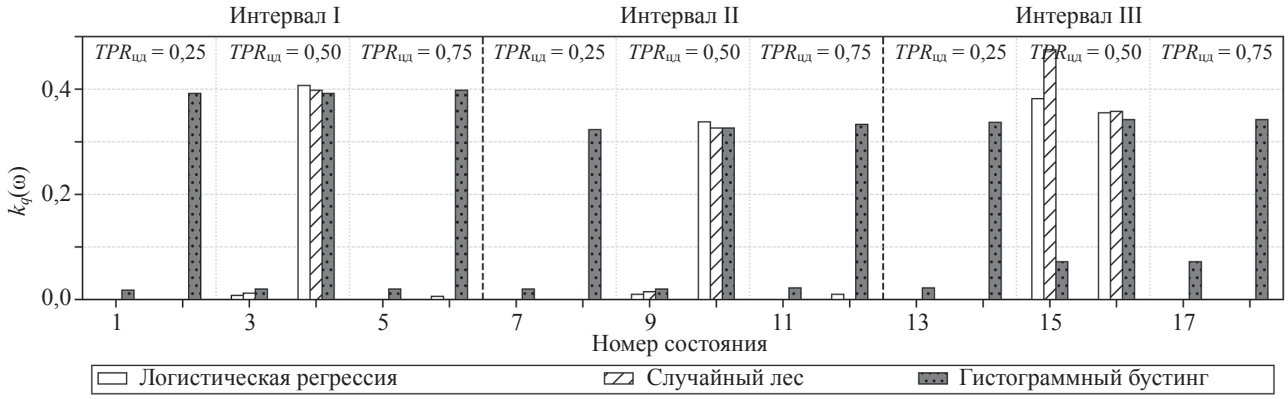


Рисунок. Значения частных показателей качества моделей для 18 состояний, заданных сочетаниями временного интервала, целевого уровня полноты ($TPR_{цд}$ — целевое значение полноты, True Positive Rate) и режима интенсивности трафика

Figure. Values of partial model quality indicators for 18 states defined by combinations of the time interval, target recall level ($TPR_{цд}$ — target True Positive Rate value), and traffic intensity mode

взвешенной суммы и взвешенного произведения [15]. При одинаковых весах показателей PPV_q, TPR_q и TNR_q равновзвешенная сумма (S_q) и геометрическое среднее (G_q) определены выражениями:

$$S_q = \sum_{j=1}^{m} w_j x_{qj} = \frac{PPV_q + TPR_q + TNR_q}{3}, \quad \sum_{j=1}^{m} w_j = 1;$$

$$G_q = PPV_{\omega}^{w_1} \times TPR_q^{w_2} \times TNR_q^{w_3} = \sqrt[3]{PPV_q \times TPR_q \times TNR_q}$$

где PPV_q, TPR_q, TNR_q — показатели модели q; w_1, w_2, w_3 — весовые коэффициенты j-го показателя (при равных весах $w_1 = w_2 = w_3 = 1/3$). Для обоих показателей предпочтительно большее значение.

Байесовская вероятность наличия атаки при отнесении записи моделью к легитимному классу определена по формуле, полученной из теоремы Байеса [16]:

$$P(A|\hat{y} = 0) = \frac{P(\hat{y} = 0|A)P(A)}{P(\hat{y} = 0|A)P(A) + P(\hat{y} = 0|\bar{A})P(\bar{A})};$$

$$B_q = P(A|\hat{y} = 0) = \frac{FNR_q \times P(A)}{FNR_q \times P(A) + TNR_q[1 - P(A)]},$$

где A — событие наличия атаки; $P(A)$ — априорная вероятность атаки, определяемая как доля вредоносных записей в исследуемой выборке; FNR_q — доля пропущенных атак; TNR_q — специфичность модели; q — индекс оцениваемой модели; B_q — вероятность того, что запись, отнесенная моделью к легитимному классу, в действительности является вредоносной. Для S_q, G_q и K_q^{int} предпочтительно большее значение, которые характеризуют качество модели; для B_q — меньшее, так как данный показатель отражает риск пропуска атаки.

Таблица 3. Сравнение подходов к оценке моделей

Table 3. Comparison of approaches to model evaluation

Подход	Учитываемые характеристики	Базовые показатели ЛР/СЛ/ГБ	Порядок предпочтения моделей	Методические свойства и ограничения
Равновзвешенная сумма показателей (KPI)	Рассчитанные по всей исследуемой выборке	0,792832/ 0,795734/ 0,791438	СЛ–ЛР–ГБ	Допускает компенсацию низкого показателя высокими значениями остальных и не учитывает различия между состояниями
Геометрическое среднее	*PPV*, *TPR*, *TNR*, рассчитанные по всей исследуемой выборке	0,723980/ 0,729243/ 0,720686	СЛ–ЛР–ГБ	Сильнее реагирует на низкое значение отдельного показателя, но не выявляет условия частного снижения качества
Байесовская оценка	*FNR*, *TNR* и доля атак ($P(A)$), рассчитанные по всей исследуемой выборке	0,441111/ 0,437817/ 0,443264	СЛ–ЛР–ГБ	Показывает вероятность атаки при отрицательном решении, но зависит от доли атак и не учитывает изменение условий
Предложенная методика	*PPV*, *TPR*, *TNR*, рассчитанные отдельно для каждого состояния, уверенность решений, время обработки и неблагоприятные отклонения	0,102880/ 0,100634/ 0,264163	ГБ–ЛР–СЛ	Выявляет снижение качества в одном из состояний функционирования, но требует предварительного задания состояний, параметров и калибровки вероятностей

Примечание. ЛР — логистическая регрессия; СЛ — случайный лес; ГБ — гистограммный бустинг.

В табл. 3 приведены значения базовых показателей и интегральной оценки, а также соответствующий порядок предпочтения моделей.

Прямое сравнение значений показателей не проводилось, поскольку они отражают различные характеристики работы модели. Равновзвешенная сумма, геометрическое среднее и байесовская оценка, рассчитанные по всей выборке при заданном пороге, обеспечивающем полноту не ниже 0,50, показали преимущество случайного леса.

В предложенной методике оценка выполнялась для нескольких порогов и условий функционирования модели. У логистической регрессии и случайного леса в отдельных состояниях получено $k_q(\omega) = 0$, что уменьшило их интегральные оценки. Для гистограммного градиентного бустинга получено наибольшее значение $K_q^{\text{int}} = 0{,}264163$. Таким образом, глобальные показатели и предложенная интегральная оценка сформировали различное ранжирование моделей. При оценке по тестовой выборке в целом преимущество получил случайный лес, а при учете всех 18 состояний — гистограммный градиентный бустинг. Полученные выводы относятся к условиям проведенного эксперимента.

Заключение

Разработана методика интегральной оценки качества моделей машинного обучения для обнаружения кибератак в сетевом трафике. Для каждого сочетания временного интервала наблюдения, целевого уровня полноты и режима интенсивности трафика рассчитывался частный показатель качества, учитывающий результаты классификации, калиброванную уверенность модели, потери от ошибочных решений и время обработки. Частные показатели объединены с учетом числа наблюдений, после чего итоговая оценка корректировалась по величине отклонений ниже средневзвешенного уровня.

В эксперименте на наборе данных GothamDataset2025 модель случайный лес показала наибольшие значения сбалансированной точности — 0,693933 и коэффициента корреляции Мэтьюса — 0,466685. Наибольшую интегральную оценку, равную 0,264163, получил гистограммный градиентный бустинг. При целевом уровне полноты 0,25 у логистической регрессии и случайного леса выявлены состояния с нулевым значением частного показателя, тогда как у гистограммного градиентного бустинга его значения оставались положительными во всех 18 состояниях. Результаты относятся к набору данных GothamDataset2025, принятой системе состояний и установленным значениям параметров методики.

Литература

1. Мещеряков Р.В., Мельников С.Ю., Пересыпкин В.А., Хорев А.А. Перспективные направления применения технологий искусственного интеллекта при защите информации // Вопросы кибербезопасности. 2024. № 4 (62). С. 2–12. doi: 10.21681/2311-3456-2024-4-02-12
2. Chicco D., Tötsch N., Jurman G. The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation // BioData Mining. 2021. V. 14. P. 13. doi: 10.1186/s13040-021-00244-z
3. Arp D., Quiring E., Pendlebury F., Warnecke A., Pierazzi F., Wressnegger C., et al. Dos and don'ts of machine learning in computer security // Proc. of the 31st USENIX Security Symposium. 2022. P. 3971–3988.
4. Lu J., Liu A., Dong F., Gu F., Gama J., Zhang G. Learning under concept drift: a review // IEEE Transactions on Knowledge and Data Engineering. 2019. V. 31. N 12. P. 2346–2363. doi: 10.1109/tkde.2018.2876857
5. Bagui S.S., Khan M.P., Valmyr C., Bagui S.C., Mink D. Model retraining upon concept drift detection in network traffic big data // Future Internet. 2025. V. 17. N 8. P. 328. doi: 10.3390/fi17080328
6. Булгакова Е.В., Дойников Д.С., Кубанков А.Н. Проблема точности и объяснимости при внедрении искусственного интеллекта в системы управления информацией и событиями безопасности // Наукоемкие технологии в космических исследованиях Земли. 2025. Т. 17. № 3. С. 35–41. doi: 10.36724/2409-5419-2025-17-3-35-41
7. Ичетовкин Е.А. Исследование устойчивости систем обнаружения вторжений с компонентами машинного обучения к состязательным атакам // Вестник Астраханского государственного технического университета. Серия: Управление, вычислительная техника и информатика. 2025. № 2. С. 76–87. doi: 10.24143/2072-9502-2025-2-76-87
8. Wang L.-H., Dai Q., Du T., Chen L.-F. Lightweight intrusion detection model based on CNN and knowledge distillation // Applied Soft Computing. 2024. V. 165. P. 112118. doi: 10.1016/j.asoc.2024.112118
9. Лившиц И.И. Влияние современных технологий искусственного интеллекта на безопасность промышленных систем автоматизации // Автоматизация в промышленности. 2025. № 6. С. 34–37.

References

1. Meshcheryakov R.V., Melnikov S.Yu., Peresypkin V.A., Khorev A.A. Promising directions for applying artificial intelligence technologies in information protection. *Voprosy Kiberbezopasnosti*, 2024, no. 4 (62), pp. 2–12. (in Russian). doi: 10.21681/2311-3456-2024-4-02-12
2. Chicco D., Tötsch N., Jurman G. The Matthews correlation coefficient (MCC) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. *BioData Mining*, 2021, vol. 14, pp. 13. doi: 10.1186/s13040-021-00244-z
3. Arp D., Quiring E., Pendlebury F., Warnecke A., Pierazzi F., Wressnegger C., et al. Dos and don'ts of machine learning in computer security. *Proc. of the 31st USENIX Security Symposium*, 2022, pp. 3971–3988.
4. Lu J., Liu A., Dong F., Gu F., Gama J., Zhang G. Learning under concept drift: a review. *IEEE Transactions on Knowledge and Data Engineering*, 2019, vol. 31, no. 12, pp. 2346–2363. doi: 10.1109/tkde.2018.2876857
5. Bagui S.S., Khan M.P., Valmyr C., Bagui S.C., Mink D. Model retraining upon concept drift detection in network traffic big data. *Future Internet*, 2025, vol. 17, no. 8, pp. 328. doi: 10.3390/fi17080328
6. Bulgakova E.V., Doynikov D.S., Kubankov A.N. The problem of accuracy and explainability when implementing artificial intelligence in security information and event management systems. *Hi-Tech & Earth Space Research*, 2025, vol. 17, no. 3, pp. 35–41. (in Russian). doi: 10.36724/2409-5419-2025-17-3-35-41
7. Ichetovkin E.A. Investigating the resistance of intrusion detection systems with machine learning components to adversarial attacks. *Vestnik of Astrakhan State Technical University. Series: Management, Computer Sciences and Informatics*, 2025, no. 2, pp. 76–87. (in Russian). doi: 10.24143/2072-9502-2025-2-76-87
8. Wang L.-H., Dai Q., Du T., Chen L.-F. Lightweight intrusion detection model based on CNN and knowledge distillation. *Applied Soft Computing*, 2024, vol. 165, pp. 112118. doi: 10.1016/j.asoc.2024.112118
9. Livshits I.I. Impact of modern artificial intelligence technologies on the security of industrial automation systems. *Avtomatizatsiya v promyshlennosti*, 2025, no. 6, pp. 34–37. (in Russian)

10. Лившиц И.И. Интеграция требований к системам менеджмента искусственного интеллекта в автомобильной промышленности // Стандарты и качество. 2025. № 12. С. 46–50. doi: 10.35400/0038-9692-2025-12-205-25
11. Krejčí J., Stoklasa J. Aggregation in the analytic hierarchy process: Why weighted geometric mean should be used instead of weighted arithmetic mean // Expert Systems with Applications. 2018. V. 114. P. 97–106. doi: 10.1016/j.eswa.2018.06.060
12. Talpini J., Sartori F., Savi M. Enhancing trustworthiness in ML-based network intrusion detection with uncertainty quantification // Journal of Reliable Intelligent Environments. 2024. V. 10. N 4. P. 501–520. doi: 10.1007/s40860-024-00238-8
13. Fishburn P.C. Mean-risk analysis with risk associated with below-target returns // American Economic Review. 1977. V. 67. N 2. P. 116–126.
14. Belarbi O., Spyridopoulos T., Anthi E., Rana O., Carnelli P., Khan A. A Device-level IoT network traffic dataset with distributed capture and non-IID characteristics // Data. 2026. V. 11. N 8. P. 207. doi: 10.3390/data11080207
15. Triantaphyllou E., Mann S.H. An examination of the effectiveness of multi-dimensional decision-making methods: A decision-making paradox // Decision Support Systems. 1989. V. 5. N 3. P. 303–312. doi: 10.1016/0167-9236(89)90037-7
16. Fischer F. Using Bayes theorem to estimate positive and negative predictive values for continuously and ordinally scaled diagnostic tests // International Journal of Methods in Psychiatric Research. 2021. V. 30. N 2. P. e1868. doi: 10.1002/mpr.1868

10. Livshits I.I. Integration of requirements to artificial intelligence management systems in the automotive industry. *Standards and Quality*, 2025, no. 12, pp. 46–50. (in Russian). doi: 10.35400/0038-9692-2025-12-205-25
11. Krejčí J., Stoklasa J. Aggregation in the analytic hierarchy process: Why weighted geometric mean should be used instead of weighted arithmetic mean. *Expert Systems with Applications*, 2018, vol. 114, pp. 97–106. doi: 10.1016/j.eswa.2018.06.060
12. Talpini J., Sartori F., Savi M. Enhancing trustworthiness in ML-based network intrusion detection with uncertainty quantification. *Journal of Reliable Intelligent Environments*, 2024, vol. 10, no. 4, pp. 501–520. doi: 10.1007/s40860-024-00238-8
13. Fishburn P.C. Mean-risk analysis with risk associated with below-target returns. *American Economic Review*, 1977, vol. 67, no. 2, pp. 116–126.
14. Belarbi O., Spyridopoulos T., Anthi E., Rana O., Carnelli P., Khan A. A Device-level IoT network traffic dataset with distributed capture and non-IID characteristics. *Data*, 2026, vol. 11, no. 8, pp. 207. doi: 10.3390/data11080207
15. Triantaphyllou E., Mann S.H. An examination of the effectiveness of multi-dimensional decision-making methods: A decision-making paradox. *Decision Support Systems*, 1989, vol. 5, no. 3, pp. 303–312. doi: 10.1016/0167-9236(89)90037-7
16. Fischer F. Using Bayes theorem to estimate positive and negative predictive values for continuously and ordinally scaled diagnostic tests. *International Journal of Methods in Psychiatric Research*, 2021, vol. 30, no. 2, pp. e1868. doi: 10.1002/mpr.1868

Авторы

Лукоянов Алексей Александрович — аспирант, Российский государственный гуманитарный университет, Москва, 125047, Российская Федерация; специалист по информационной безопасности, АНО «Проектный офис по развитию туризма и гостеприимства Москвы», Москва, 125009, Российская Федерация, https://orcid.org/0009-0001-4519-6882, 9261781233@mail.ru

Рогозин Евгений Алексеевич — доктор технических наук, профессор, профессор, Воронежский институт МВД России, Воронеж, 394065, Российская Федерация, SC 55579562100, https://orcid.org/0000-0002-4455-7535, evgenirogozin@yandex.ru

Рогозин Руслан Евгеньевич — кандидат технических наук, доцент, МИРЭА — Российский технологический университет, Москва, 119454, Российская Федерация, https://orcid.org/0000-0002-2913-6038, ruslan-96-01-09@mail.ru

Лившиц Илья Иосифович — доктор технических наук, доцент, профессор практики, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, SC 57191569306, https://orcid.org/0000-0003-0651-8591, Livshitz.il@yandex.ru

Authors

Aleksey A. Lukoyanov — PhD Student, Russian State University for the Humanities (RSUH), Moscow, 125047, Russian Federation; Information Security Specialist, Autonomous Non-Profit Organization “Project Office for Tourism and Hospitality Development of Moscow”, Moscow, 125009, Russian Federation, https://orcid.org/0009-0001-4519-6882, 9261781233@mail.ru

Evgeny A. Rogozin — D.Sc., Full Professor, Voronezh Institute of the Ministry of Internal Affairs of Russia, Voronezh, 394065, Russian Federation, SC 55579562100, https://orcid.org/0000-0002-4455-7535, evgenirogozin@yandex.ru

Ruslan E. Rogozin — PhD, Associate Professor, MIREA — Russian Technological University, Moscow, 119454, Russian Federation, https://orcid.org/0000-0002-2913-6038, ruslan-96-01-09@mail.ru

Ilya I. Livshits — D.Sc., Associate Professor, Professor of Practice, ITMO University, Saint Petersburg, 197101, Russian Federation, SC 57191569306, https://orcid.org/0000-0003-0651-8591, Livshitz.il@yandex.ru

Статья поступила в редакцию 29.04.2026
Одобрена после рецензирования 20.08.2026
Принята к печати 23.09.2026

Received 29.04.2026
Approved after reviewing 20.08.2026
Accepted 23.09.2026

Работа доступна по лицензии
Creative Commons
«Attribution-NonCommercial»