

ИНТЕРПРЕТАЦИОННАЯ МОДЕЛЬ НАБЛЮДЕНИЯ И ДИАГНОСТИРОВАНИЯ СОСТОЯНИЯ ЭЛЕМЕНТОВ БЛОКЧЕЙН-СИСТЕМ ПО РЕШЕНИЯМ МОДЕЛЕЙ МАШИННОГО ОБУЧЕНИЯ

Лисов Д.Н.¹

Ключевые слова: объяснимый искусственный интеллект, локальные атрибуции признаков, бинарная классификация угроз, качество объяснения, диагностический профиль, доверие к прогнозу, подозрительная активность, аналитическая проверка, новые угрозы безопасности, сетевые вторжения.

Аннотация

Цель работы: разработать интерпретационную модель наблюдения и диагностирования состояния элементов блокчейн-систем при появлении угроз нового типа путем совместного анализа прогнозов моделей машинного обучения и результатов их локальной интерпретации.

Методы исследования: выполнена функциональная декомпозиция наблюдения, включающего получение прогноза, построение локального объяснения, количественную оценку его свойств и формирование диагностического заключения. Для количественной оценки локального объяснения использованы две метрики качества: верность и устойчивость. Метрика верности показывает, насколько замена выделенных объяснителем признаков изменяет выход модели по сравнению с заменой случайного набора, а метрика устойчивости характеризует сохранение состава значимых признаков у близких объектов с сопоставимым прогнозом. Значения метрик определялись путем контролируемой замены входных признаков и сопоставления локальных объяснений, а их диагностическая информативность оценивалась с помощью характеристических кривых. При экспериментальном воспроизведении угрозы нового типа один тип атаки исключался из обучающих данных и сохранялся в тестовой выборке, после чего проверялись методы объяснения и их сочетания.

Результаты исследования: предложена интерпретационная модель, в которой результат локального объяснения включается в наблюдение за решением ML-модели и формирование диагностического состояния объекта. Измеримые свойства объяснения преобразуются в вектор показателей, по которому определяется уровень доверия. Если диагностический профиль объяснения не соответствует характерным профилям известных корректных решений, объект получает промежуточное состояние «подозрительное» и сохраняется для дополнительной проверки. Экспериментальное исследование показало, что показатели качества объяснения содержат диагностическую информацию, не представленную в выходном классе модели, однако ее выраженность зависит от данных и применяемого объяснителя. При поступлении активности, отсутствовавшей в обучающих данных, интерпретационный контур позволил сохранить в области наблюдения часть объектов, пропущенных базовой моделью. Совместное использование нескольких методов объяснения оказалось полезным только при взаимном дополнении получаемых атрибуций и не обеспечило универсального преимущества. Полученные результаты подтверждают возможность выделять решения, требующие дополнительной проверки при появлении ранее не наблюдавшихся угроз. Базовый классификатор при этом не изменяется, а подозрительное состояние не отождествляется с атакой.

DOI: 10.24412/1994-1404-2026-3-88-99

Введение и постановка задачи

Национальные блокчейн-системы и платформы экономики данных представляют собой распределенные цифровые инфраструктуры, объединяющие узлы хранения и обработки данных, сетевые

и интеграционные сервисы, средства удостоверения, механизмы исполнения транзакций, смарт-контракты и прикладные приложения. Взаимозависимость перечисленных компонентов приводит к тому, что локаль-

¹ Лисов Дмитрий Николаевич, аспирант, младший научный сотрудник центра информационных технологий и искусственного интеллекта АНОО ВО «Научно-технологический университет «Сириус», пгт Сириус, Краснодарский край, Российская Федерация. ORCID: 0000-0002-8214-4109.

E-mail: lisov.dn@talantiuspeh.ru

ное нарушение способно повлиять на выполнение функций всей системы. Поэтому ее защищенность определяется не только стойкостью отдельных механизмов безопасности, но и способностью продолжать функционирование при изменении условий и характера деструктивных воздействий [1]. Необходимость обеспечения такой способности обоснована в концепции киберустойчивости национальных цифровых платформ и блокчейн-экосистем [2], а возможные направления воздействия на элементы блокчейн-инфраструктуры конкретизированы в модели квантовых угроз [3]. Систематический учет изменения угроз и появления новых сценариев их реализации предусмотрен также Методикой оценки угроз безопасности информации ФСТЭК России². В этих условиях противодействие угрозам нового типа требует не только совершенствования предотвращающих мер, но и развития средств наблюдения, позволяющих обнаруживать изменения состояния элементов системы при неполноте предварительных сведений о возможном воздействии.

Источниками сведений о состоянии блокчейн-системы могут служить сетевой трафик, параметры транзакций, обращения к программным интерфейсам, события интеграционных сервисов и результаты выполнения прикладных функций. В исследованиях эти данные используются при разработке средств обнаружения аномалий, мошенничества и вредоносной активности на основе машинного обучения. Основные направления объединения блокчейн-технологий и ML систематизированы в работе [4], а в исследовании [5] классификаторы применены для выявления аномальных транзакций с последующим объяснением полученных решений. ML-модель в таких разработках выступает одним из возможных средств распознавания угроз, применимость которого определяется архитектурой и назначением конкретного элемента. Поэтому настоящее исследование ограничено теми элементами блокчейн-системы, для которых решение задачи распознавания угроз предусматривает использование ML-модели.

При включении ML-модели в контур, связанный с безопасностью, возникает вопрос о контролируемости сформированных ею решений [6]. Выходной класс показывает результат распознавания, но не раскрывает, какие свойства объекта определили этот результат и соответствует ли поведение модели изученным закономерностям. Методы объяснимого искусственного интеллекта частично устраняют эту неопределенность и позволяют установить влияние отдельных признаков на прогноз [7, 8]. В существующих системах обнаружения вторжений объяснение преимущественно сопровождает уже полученное решение. В XAI-XGBoost методы SHAP и LIME используются для глобального и локального анализа прогнозов модели [9]. В рабо-

те [10] угрозы нулевого дня распознаются специализированной ML-моделью, а LIME и SHAP применяются для последующего анализа результата. Аналогичная последовательность используется в системе X-CBA [11]. Рассмотренные решения повышают прозрачность обнаружения и позволяют специалисту проверить содержательные основания прогноза, однако результат интерпретации не участвует в определении состояния объекта и не показывает, насколько свойства полученного объяснения согласуются с решением классификатора.

Попытки включить XAI непосредственно в дальнейшую обработку данных привели к появлению более сложных архитектур. В работе [12] атрибуции первой модели поступают на вход дополнительному классификатору, благодаря чему выявленные XAI зависимости используются при повторном распознавании DDoS-атак. Другой способ операционного применения объяснений предложен в работе [13]. В разработанной авторами федеративной системе SHAP обеспечивает прослеживаемость решений агентов, используется для выявления нестабильного поведения и участвует в ограничении влияния потенциально скомпрометированных узлов на обновление общей модели. Обе работы подтверждают, что результаты интерпретации могут применяться не только для представления прогноза. Вместе с тем в первом случае создается дополнительный обучаемый детектор, а во втором изменяется политика федеративного обновления. Следовательно, XAI встраивается в механизм обучения или адаптации, тогда как задача независимого наблюдения за уже сформированным прогнозом остается нерешенной.

Наиболее близкий к указанной задаче подход предложен Mougan и соавторами [14]. Авторы исходят из того, что изменение поступающих данных может проявляться не только в распределении входных признаков или прогнозов, но и в распределении объяснений. Атрибуции, рассчитанные для исходной и новой выборки, передаются бинарному классификатору, результат которого характеризует наличие сдвига объяснений. XAI здесь уже участвует в формировании показателя мониторинга, а не только сопровождает готовый прогноз. Тем не менее единицей наблюдения остается совокупность новых данных: подход устанавливает факт изменения взаимодействия модели с потоком объектов, но не формирует диагностическое заключение для каждого объекта, не оценивает качество его локального объяснения и не связывает полученную оценку с исходным классом. Смежная работа [15] посвящена оценке изменения производительности ML-модели и установлению входных признаков, связанных с предполагаемым ухудшением ее качества. XAI в этом случае объясняет результат мониторинга и помогает определить возможные причины деградации, однако также не используется для диагностирования отдельного объекта. Таким образом, существующие исследования содержат отдельные предпосылки интерпретационного наблюдения — локальное объяснение

² Методический документ. Методика оценки угроз безопасности информации (утв. ФСТЭК России 05.02.2021). Пп. 2.4, 2.14 // СПС «КонсультантПлюс».

прогноза, использование атрибуций при дополнительном распознавании и обнаружение сдвига объяснений, но не устанавливают единой связи между прогнозом конкретного объекта, измеримыми свойствами его объяснения и диагностическим состоянием.

Для построения такой связи недостаточно использовать непосредственно значения атрибуций. Они характеризуют вклад признаков, но не позволяют установить, действительно ли выделенные признаки определяют поведение модели и сохраняется ли полученная структура объяснения относительно сопоставимых объектов. Необходимость количественной оценки XAI обоснована в систематическом обзоре [16], где показано, что качество объяснения оценивается с помощью групп метрик, характеризующих его верность, устойчивость, сложность и локализацию. Инструментарий Quantus предоставляет унифицированные процедуры вычисления соответствующих показателей [17], однако их значения зависят от модели, данных, объяснителя и способа проверки, что подтверждается критическим анализом современного состояния XAI [18]. В существующих исследованиях такие показатели преимущественно используются для сравнения объяснителей на выборке. Для наблюдения за состоянием объекта требуется иное применение: свойства объяснения должны оцениваться для конкретного прогноза, сохраняться в виде отдельных диагностических признаков и рассматриваться совместно с исходным выходом модели. Поскольку разные XAI-методы используют неодинаковые способы определения вклада признаков, сформированные ими объяснения и оценки их свойств могут различаться для одного объекта. Это различие создает основание для проверки как отдельных объяснителей, так и их совместного использования. При этом низкое качество объяснения не доказывает ошибочность прогноза и потому не должно приводить к его автоматической инверсии.

Научная проблема состоит в отсутствии формальной связи между бинарным прогнозом ML-модели, количественными свойствами его локального объяснения и диагностическим состоянием наблюдаемого элемента блокчейн-системы. Вследствие этого объяснение либо предъявляется после принятия решения, либо используется для изменения обучаемого механизма, но не обеспечивает независимое диагностирование конкретного объекта в рабочем цикле наблюдения. Гипотеза исследования заключается в том, что локальное объяснение содержит измеримые свойства решения классификатора, не представленные в его бинарном выходе, а их совместное рассмотрение с прогнозом позволяет выделять объекты, для которых принятое решение не получает достаточного интерпретационного подтверждения. Предполагается, что такие отклонения могут сопровождать появление ранее не наблюдавшихся угроз, однако сами по себе не являются доказательством атаки и требуют отдельного диагностического состояния. Дополнительно проверяется частное предположение о том, что совместное использование

результатов нескольких XAI-методов может повысить диагностическую информативность описания по сравнению с отдельным объяснителем за счет учета различий в определении значимости признаков.

Цель работы состоит в разработке интерпретационной модели наблюдения и диагностирования состояния элементов блокчейн-систем, для которых распознавание угроз выполняется с применением ML-моделей. Для достижения цели необходимо определить состав результатов ML и XAI анализа, доступных при наблюдении за отдельным объектом, установить способ количественного представления свойств локального объяснения, формализовать их использование совместно с решением классификатора при формировании диагностического состояния, определить связь состояния с сигналом и объяснимым отчетом, а также экспериментально оценить поведение модели при поступлении активности, отсутствовавшей в обучающих данных.

Научная новизна состоит в разработке модели интерпретационного наблюдения, в которой количественные свойства локального объяснения используются как диагностические признаки отдельного объекта. В отличие от постфактум-интерпретации предлагаемая модель учитывает не только содержание объяснения, но и степень подтверждения им прогноза. В отличие от решений, передающих атрибуции дополнительному классификатору, исходная ML-модель не заменяется и не переобучается. В отличие от мониторинга сдвига объяснений диагностический результат формируется для каждого объекта, а недостаточное интерпретационное подтверждение представляется самостоятельным состоянием, требующим аналитической проверки.

Интерпретационная модель наблюдения и диагностирования состояния элемента блокчейн-системы

Предлагаемая модель строится вокруг среды национального блокчейна. Под такой средой понимается не только распределенный реестр, но и связанные с ним сетевые узлы, валидаторы, сервисы доступа, средства обработки транзакций, механизмы исполнения смарт-контрактов, прикладные модули и интерфейсы взаимодействия с внешними системами³. Обозначим совокупность наблюдаемых элементов как $B = \{B_1, \dots, B_n\}$. Данные об их работе поступают последовательно, поэтому введем множество моментов времени $\mathcal{T} = \{t_1, \dots, t_\delta\}$. Для каждого элемента B_i в момент $t \in \mathcal{T}$ формируется вектор входных данных $X_{(B_i)}(t) = (x_1, x_2, \dots, x_k) \in \mathbb{R}^k$. В зависимости от назначения элемента этот вектор может описывать сетевое поведение, параметры транзакции, обращение

³ Recommendation ITU-T X.1413 (05/2025). Security controls for distributed ledger technology. URL: <https://www.itu.int/rec/T-REC-X.1413/en> (дата обращения: 08.08.2026)

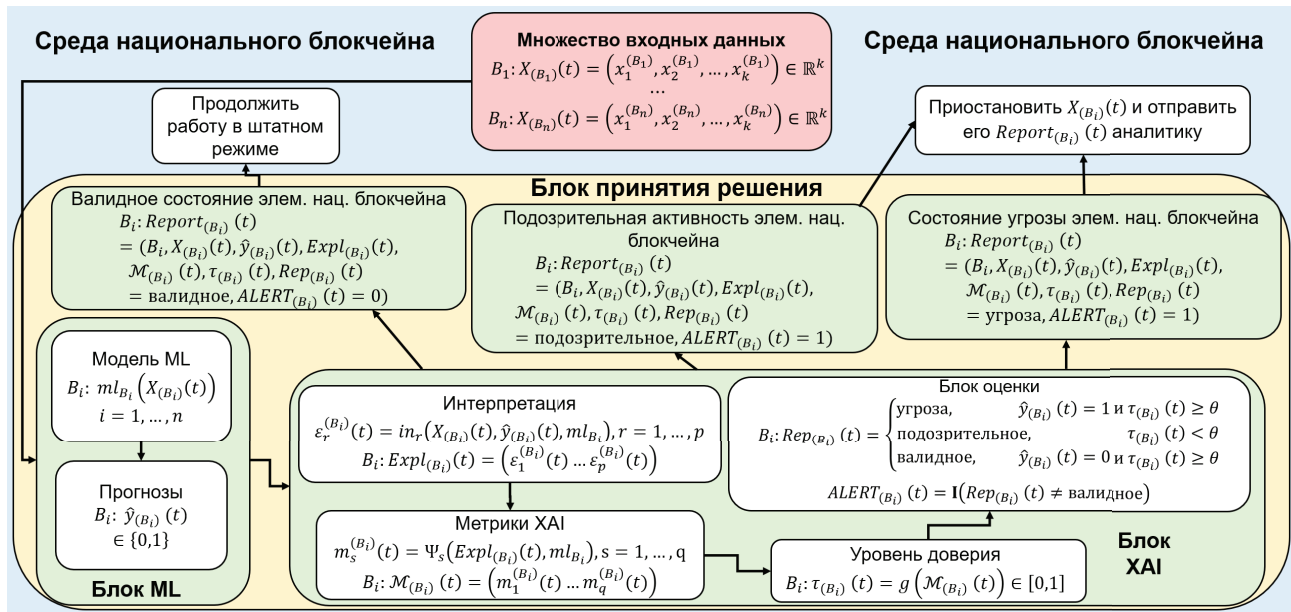


Рис. 1. Интерпретационная модель наблюдения и диагностирования состояния элемента блокчейн-системы

к программному интерфейсу, событие интеграционного сервиса, выполнение смарт-контракта или другую наблюдаемую активность.

Такая постановка имеет два следствия. Во-первых, модель сохраняет применимость на разных уровнях блокчейн-платформы. Во-вторых, наблюдение не сводится только к анализу финансовых транзакций. Оно может быть организовано для любого элемента, состояние которого влияет на безопасность, доверие, непрерывность работы или корректность предоставляемых сервисов. При этом состав входных признаков и применяемые средства машинного обучения могут различаться: сетевой узел, механизм обработки транзакций и прикладной сервис решают разные задачи и оставляют разные наблюдаемые следы. На рис. 1 представлена общая схема интерпретационной модели наблюдения и диагностирования состояния элемента блокчейн-системы.

Как видно из рис. 1, обработка начинается с получения прогноза ML-модели. Интерпретационный контур не заменяет этот прогноз и не создает вместо исходной модели новый классификатор. Его задача состоит в том, чтобы дополнить полученный ответ сведениями о том, какие признаки определили решение и насколько свойства сформированного объяснения соответствуют условиям его дальнейшего использования.

Пусть $ML = \{ml_1, \dots, ml_n\}$ — множество обученных бинарных моделей машинного обучения, применяемых для наблюдаемых элементов. Каждому элементу B_i сопоставляется модель ml_{B_i} , которая получает вектор $X_{(B_i)}(t)$ и формирует прогноз:

$$\hat{Y}_{(B_i)}(t) = ml_{B_i}(X_{(B_i)}(t)) \in \{0, 1\}, \quad (1)$$

где значение 0 соответствует прогнозу нормы, значение 1 — прогнозу угрозы. Здесь важно различать прогноз и фактическое состояние объекта. Бинарная модель должна выбрать один из предусмотренных

классов даже тогда, когда поступившие данные отличаются от наблюдавшихся при ее обучении. Поэтому значение $\hat{Y}_{(B_i)}(t)$ показывает решение классификатора, но само по себе не позволяет установить, насколько уверенно это решение подтверждается особенностями конкретного объекта.

Дополнительные сведения о сформированном прогнозе получают с помощью локальной интерпретации. Для рассматриваемого элемента задается конфигурация $In = \{in_1, \dots, in_p\}$, совместимая с его ML-моделью и представлением входных данных. При $p = 1$ она включает один XAI-метод, а при $p > 1$ может объединять несколько совместимых методов, описывающих один и тот же прогноз. Конкретный состав конфигурации и способ совместного представления ее результатов зависят от реализации интерпретационного контура. Каждый метод in_r применяется к объекту $X_{(B_i)}(t)$, прогнозу $\hat{Y}_{(B_i)}(t)$ и соответствующей модели ml_{B_i} :

$$\varepsilon_r^{(B_i)}(t) = in_r(X_{(B_i)}(t), \hat{Y}_{(B_i)}(t), ml_{B_i}), \quad r = 1, \dots, p. \quad (2)$$

Результат $\varepsilon_r^{(B_i)}(t)$ представляет локальное объяснение прогноза и характеризует направление и относительную силу влияния входных признаков на решение ML-модели. Результаты выбранной XAI-конфигурации образуют набор $Expl_{(B_i)}(t) = (\varepsilon_1^{(B_i)}(t), \dots, \varepsilon_p^{(B_i)}(t))$.

Таким образом, для каждого наблюдаемого элемента используется подходящее ему средство локальной интерпретации либо выбранная совместимая комбинация.

На следующем этапе осуществляется переход от результата локальной интерпретации к измеряемым показателям качества объяснения. Атрибуции показывают, какие признаки и в каком направлении повлияли

на прогноз, однако сами по себе не позволяют количественно охарактеризовать свойства полученного объяснения. Поэтому в модель вводится множество метрик $\Psi = \{\Psi_1, \dots, \Psi_q\}$. Каждая метрика Ψ_s оценивает отдельное свойство объяснения с учетом поведения соответствующей ML-модели и возвращает значение $m_s^{(B_i)}(t)$:

$$m_s^{(B_i)}(t) = \Psi_s(Expl_{(B_i)}(t), ml_{B_i}), s = 1, \dots, q. \quad (3)$$

В зависимости от рассматриваемого элемента и выбранной ХАИ-конфигурации вектор может включать метрики верности, устойчивости, согласованности и другие характеристики локального объяснения, для которых определены воспроизводимые процедуры расчета. При этом метрики не заменяют атрибуты и не формируют новый класс объекта. Они переводят свойства объяснения в измеряемую форму, необходимую для их последующего совместного учета.

Полученные значения метрик образуют вектор показателей качества объяснения

$$\mathcal{M}_{(B_i)}(t) = (m_1^{(B_i)}(t), \dots, m_q^{(B_i)}(t)), \text{ который пред-}$$

ставляет количественное описание свойств объяснения, сформированного для текущего прогноза. Каждая его компонента сохраняет смысл соответствующей метрики, а весь вектор используется как единое диагностическое описание свойств объяснения при дальнейшем формировании уровня доверия.

Для совместного учета рассчитанных показателей вводится правило g , преобразующее вектор $\mathcal{M}_{(B_i)}(t)$ в скалярный уровень доверия:

$$\tau_{(B_i)}(t) = g(\mathcal{M}_{(B_i)}(t)) \in [0,1]. \quad (4)$$

Правило g устанавливает связь между совокупностью измеренных свойств объяснения и итоговой оценкой. Модель не ограничивает его одной заранее заданной математической операцией, поскольку характер связи зависит от наблюдаемого элемента, ML-модели и используемой ХАИ-конфигурации. При этом $\tau_{(B_i)}(t)$ не является вероятностью угрозы и не заменяет прогноз $\hat{y}_{(B_i)}(t)$. Эти значения описывают разные стороны рассматриваемого объекта: прогноз указывает выбранный ML-моделью класс, а уровень доверия характеризует свойства объяснения этого прогноза.

Для перехода от прогноза и уровня доверия к диагностическому состоянию вводится порог $\theta \in [0,1]$. Состояние наблюдаемого элемента определяется выражением:

$$Rep_{(B_i)}(t) = \begin{cases} \text{угроза, } \hat{y}_{(B_i)}(t) = 1 \text{ и } \tau_{(B_i)}(t) \geq \theta, \\ \text{подозрительное, } \tau_{(B_i)}(t) < \theta, \\ \text{валидное, } \hat{y}_{(B_i)}(t) = 0 \text{ и } \tau_{(B_i)}(t) \geq \theta. \end{cases} \quad (5)$$

Значение $Rep_{(B_i)}(t)$ является диагностическим заключением интерпретационного контура и не рассматривается как достоверно известная фактическая метка объекта.

Состояние «угроза» формируется при совпадении двух условий: ML-модель обнаружила признаки угрозы, а свойства объяснения соответствуют установленному уровню доверия. Такое заключение сохраняет исходный прогноз модели и одновременно дополняет его результатами интерпретационного контроля.

«Подозрительное состояние» возникает при недостаточном уровне доверия независимо от значения бинарного прогноза. В данном случае прогноз не отменяется и не заменяется противоположным, однако его автоматическое применение ограничивается. Тем самым состояние отражает недостаточность основания для однозначного отнесения объекта к валидному состоянию или угрозе.

«Валидное состояние» соответствует прогнозу нормы, объяснение которого удовлетворяет установленному условию. Такой объект может продолжить обработку в штатном режиме, поскольку интерпретационный контур не выявил оснований для дополнительной проверки.

Для передачи результата во внешний контур вводится бинарный сигнал:

$$ALERT_{(B_i)}(t) = I(Rep_{(B_i)}(t) \neq \text{валидное}), \quad (6)$$

где $I(\cdot)$ обозначает индикатор выполнения условия. Значение $ALERT = 1$ означает необходимость передачи результата во внешний контур, а основание формирования сигнала сохраняется в Rep .

Итогом обработки объекта является отчет:

$$Report_{(B_i)}(t) = (B_i, X_{(B_i)}(t), \hat{y}_{(B_i)}(t), Expl_{(B_i)}(t), \mathcal{M}_{(B_i)}(t), \tau_{(B_i)}(t), Rep_{(B_i)}(t), ALERT_{(B_i)}(t)) \quad (7)$$

Отчет последовательно объединяет исходные сведения об объекте, прогноз ML-модели, результаты локальной интерпретации, значения метрик, уровень доверия и сформированное диагностическое состояние. Такое представление позволяет не только передать сигнал о необходимости реагирования, но и сохранить сведения, объясняющие причину принятого решения.

Поскольку модель предназначена для применения в действующей системе защиты, интерпретация и оценка ее результатов должны выполняться в пределах установленных ограничений:

$$T_{Expl_{(B_i)}} + T_{\tau_{(B_i)}} \leq T_{max}, \quad (8)$$

$$R(Expl_{(B_i)}) + R(\tau_{(B_i)}) \leq R_{max}. \quad (9)$$

Здесь $T_{Expl_{(B_i)}}$ обозначает время получения локального объяснения, а $T_{\tau_{(B_i)}}$ — время расчета метрик и уровня доверия. Аналогично $R(\cdot)$ характеризует используемые вычислительные ресурсы.

Таким образом, модель формализует включение результатов локальной интерпретации в наблюдение за решением ML-модели и диагностирование состояния поступившего объекта. Свойства объяснения

представляются вектором $\mathcal{M}$, из которого правило g формирует уровень доверия τ . Совместный учет решения $\hat{y}$ и уровня доверия определяет состояние Rep , на основе которого формируются сигнал $ALERT$ и отчет $Report$. Такая последовательность позволяет использовать локальное объяснение не только для последующего ознакомления аналитика, но и как источник сведений при формировании диагностического результата. Конкретные способы вычисления метрик, построения правила g и настройки порога θ моделью не фиксируются.

Экспериментальное исследование интерпретационной модели

Экспериментальное исследование проведено на наборе сетевого трафика CICIDS2017 [19]⁴. Для воспроизведения поступления данных, отсутствовавших при обучении ML-модели, записи с выбранным видом вредоносной активности полностью исключались из обучающей части и сохранялись в тестовой. Исследование включало три сценария: DDoS, PortScan и совместное появление DoS GoldenEye, FTP-Patator и SSH-Patator. Остальные данные разделялись на обучающую, эталонную и тестовую части. Обучающая часть использовалась для построения бинарной ML-модели, эталонная — для расчета распределений показателей объяснения и выбора порога, тестовая — для итогового сравнения исходного решения модели с результатом интерпретационного контроля. Во всех сценариях сохранялись одинаковые предобработка и структура классификатора.

Для каждого объекта вычислялись прогноз ML-модели и локальные атрибуции признаков методами Integrated Gradients, Feature Ablation, DeepLift, GradientSHAP и LIME. Рассматривались одиночные методы, их попарные сочетания и общая конфигурация из пяти объяснителей. В конфигурациях, объединяющих несколько объяснителей, абсолютные атрибуции каждого метода отдельно приводились к диапазону от 0 до 1 относительно 95-го перцентиля значений текущего объекта, после чего соответствующие оценки признаков усреднялись. Такое преобразование обеспечивало сопоставимость результатов методов, исходные масштабы атрибуций которых различаются.

Диагностический профиль объяснения формировался из показателей верности и устойчивости. Показатель верности характеризует, насколько изменение наиболее значимых по объяснению признаков воздействует на выход модели сильнее, чем изменение случайно выбранных признаков того же количества:

$$\Delta_{\text{знач}}(X_{(B_i)}(t)) = \left| z(X_{(B_i)}(t)) - z(\tilde{X}_{\text{знач}}^{(B_i)}(t)) \right|, \quad (10)$$

$$\Delta_{\text{контр}}(X_{(B_i)}(t)) = \frac{1}{5} \sum_{q=1}^5 \left| z(X_{(B_i)}(t)) - z(\tilde{X}_{\text{контр}, q}^{(B_i)}(t)) \right|, \quad (11)$$

$$m_{\text{верн}}(X_{(B_i)}(t)) = \frac{\Delta_{\text{знач}}(X_{(B_i)}(t))}{\Delta_{\text{знач}}(X_{(B_i)}(t)) + \Delta_{\text{контр}}(X_{(B_i)}(t)) + \varepsilon}, \quad (12)$$

где $z(X_{(B_i)}(t))$ обозначает выход базовой модели, а $\tilde{X}_{\text{знач}}^{(B_i)}$ и $\tilde{X}_{\text{контр}, q}^{(B_i)}$ представляют измененные варианты

анализируемого объекта. Для формирования $\tilde{X}_{\text{знач}}^{(B_i)}$ выбирались 30% признаков с наибольшими абсолютными атрибуциями, но не менее пяти признаков. Их исходные значения заменялись медианами соответствующих признаков эталонной выборки. Контрольное воздействие рассчитывалось аналогичным образом для пяти случайных наборов той же мощности. Вызванные ими изменения выхода модели усреднялись и сопоставлялись с результатом замены значимых признаков. Поэтому показатель верности отражает, насколько признаки, выделенные объяснителем, влияют на решение модели сильнее случайного набора.

Показатель устойчивости характеризует сохранение состава значимых признаков в локальной окрестности анализируемого объекта. Для этого в эталонной выборке находился ближайший по евклидову расстоянию объект с тем же прогнозируемым классом, после чего множества значимых признаков сопоставлялись с помощью коэффициента Жаккара. Полученное сходство учитывалось при различии прогнозируемых вероятностей не более 0,15. В противном случае показатель устойчивости принимал нулевое значение. Тем самым сравнивались объекты, близкие не только в пространстве входных признаков, но и по выходу ML-модели.

$$m_{\text{уст}}(X_{(B_i)}(t)) = \frac{\left| \mathbb{T}_{(B_i)}(t) \cap \mathbb{T}_{\text{ближ}}^{(B_i)}(t) \right|}{\left| \mathbb{T}_{(B_i)}(t) \cup \mathbb{T}_{\text{ближ}}^{(B_i)}(t) \right|} \cdot \mathbf{I} \left(\left| z(X_{(B_i)}(t)) - z(X_{\text{ближ}}^{(B_i)}(t)) \right| \leq 0.15 \right), \quad (13)$$

где $\mathbb{T}_{(B_i)}(t)$ — маска значимых признаков анализируемого объекта, а $\mathbb{T}_{\text{ближ}}^{(B_i)}(t)$ — маска ближайшего сопоставимого объекта из эталонной выборки известных состояний. $\mathbf{I}(\cdot)$ принимает значение 1 при выполнении указанного условия и 0 в противном случае.

Полученные значения образуют вектор показателей качества объяснения $\mathcal{M}_{(B_i)}(t) = (m_{\text{верн}}^{(B_i)}(t), m_{\text{уст}}^{(B_i)}(t))$. В эксперимен-

тальной реализации уровень доверия $\tau_{(B_i)}$ определялся как среднее арифметическое значений метрик верности и устойчивости:

⁴ Canadian Institute for Cybersecurity. Intrusion Detection Evaluation Dataset (CIC-IDS2017). URL: <https://www.unb.ca/cic/datasets/ids-2017.html> (дата обращения: 08.08.2026).

Таблица 1

Показатели бинарного распознавания

Целевой сценарий	Accuracy	Balanced accuracy	Precision	Recall	F1
DDoS (базовая модель)	0,889	0,926	0,995	0,870	0,929
DDoS (интерпретационный контур)	0,915—0,939	0,821—0,873	0,934—0,961	0,936—0,993	0,948—0,965
PortScan (базовая модель)	0,700	0,817	0,997	0,641	0,781
PortScan (интерпретационный контур)	0,695—0,769	0,693—0,757	0,916—0,950	0,683—0,788	0,789—0,850
DoS GoldenEye + FTP-Patator + SSH-Patator (базовая модель)	0,790	0,871	0,998	0,7514	0,857
DoS GoldenEye + FTP-Patator + SSH-Patator (интерпретационный контур)	0,800—0,842	0,739—0,798	0,924—0,952	0,805—0,878	0,873—0,903

$$\tau_{(B_i)}(t) = g\left(\mathcal{M}_{(B_i)}(t)\right) = \frac{m_{\text{верн}}^{(B_i)}(t) + m_{\text{уст}}^{(B_i)}(t)}{2}. \quad (14)$$

Порог θ определялся отдельно для каждой ХАИ-конфигурации по корректно классифицированным объектам эталонной выборки и соответствовал 10-му процентилу эмпирического распределения τ . Значения ниже этой границы указывали на нетипичный для известных корректных решений профиль объяснения и приводили к формированию подозрительного состояния.

Для каждого целевого сценария исследовались пять одиночных ХАИ-методов, десять их попарных сочетаний и общий ансамбль из пяти объяснителей. Поскольку введение подозрительного состояния изменяет итоговый сигнал, полученные варианты сначала сопоставлялись с базовой ML-моделью по общим показателям бинарного распознавания. Диапазоны значений по ХАИ-конфигурациям приведены в табл. 1.

Результаты показывают, что интерпретационный контур меняет не только число правильных решений, но и соотношение разных видов ошибок. Объекты атак, пропущенные базовой моделью, частично переходят в подозрительное состояние, вследствие чего повышаются Recall и F1. Одновременно в дополнительную проверку попадает часть объектов нормы, что отражается на Precision и Balanced accuracy. Такое изменение нельзя однозначно характеризовать как улучшение или ухудшение классификации. Оно соответствует переходу от полностью автоматического бинарного решения к диагностированию случаев, в которых прогноз недостаточно подтверждается свойствами локального объяснения.

Стандартные показатели не позволяют установить, связано ли формирование подозрительного состояния с содержательными различиями объяснений. Поэтому отдельно исследовалась способность метрик верности и устойчивости, а также их совместной оценки различать правильные и ошибочные решения ML-модели. Чтобы исключить случайный характер обнаруженных различий, значения показателей сравнивались с по-

мощью критерия Манна — Уитни. При проверке нескольких ХАИ-конфигураций применялась поправка Бенджамини — Хохберга, уменьшающая вероятность ложного вывода о наличии различий. Диагностическая информативность показателей представлена в табл. 2.

Таблица 2

Диагностическая информативность показателей объяснения

Целевой сценарий	AUC верности	AUC устойчивости	AUC итоговой оценки τ
DDoS	0,715—0,901	0,479—0,896	0,769—0,932
PortScan	0,485—0,558	0,545—0,765	0,543—0,710
DoS GoldenEye + FTP-Patator + SSH-Patator	0,466—0,542	0,567—0,903	0,532—0,800

Полученные результаты подтверждают, что метрики верности и устойчивости характеризуют разные стороны локального объяснения. Метрика верности оказалась наиболее информативной в сценарии DDoS, но слабее разделяла правильные и ошибочные решения в двух других сценариях. Метрика устойчивости, напротив, сохраняла диагностическую ценность для части конфигураций PortScan и совместного сценария. Следовательно, ни одну из этих метрик нельзя считать универсальной. Их совместная оценка позволяет учитывать как влияние значимых признаков на прогноз, так и сохранение состава этих признаков в локальной окрестности объекта.

Для всех проверенных ХАИ-конфигураций различия итоговой оценки между правильными и ошибочными решениями сохранили статистическую значимость после поправки на множественные сравнения. Вместе с тем заметный разброс AUC показывает, что практическая сила такого разделения зависит от входных данных и объяснителя. Тем самым гипотеза об информативности измеримых свойств объяснения получила

Таблица 3

Охват целевой активности сигналом и дополнительная аналитическая проверка

Целевой сценарий	Охват целевой активности	Дополнительная доля на проверке	Доля нормы на проверке
DDoS (базовая модель)	0,630	0,000	0,000
DDoS (интерпретационный контур)	0,822—0,994	0,083—0,157	0,174—0,334
PortScan (базовая модель)	0,0014	0,000	0,000
PortScan (интерпретационный контур)	0,110—0,434	0,069—0,176	0,178—0,324
DoS GoldenEye + FTP-Patator + SSH-Patator (базовая модель)	0,289	0,000	0,000
DoS GoldenEye + FTP-Patator + SSH-Patator (интерпретационный контур)	0,452—0,683	0,077—0,163	0,204—0,360

подтверждение, однако результаты не дают оснований рассматривать отдельную метрику или ХАИ-метод как одинаково эффективные во всех условиях.

Установленная различимость профилей объяснения показывает, что они содержат сведения, дополняющие выходной класс ML-модели. Практический эффект этой информации оценивался путем сопоставления исходного режима и режима с интерпретационным контуром. Пусть $\Omega_{\text{цел}}$ обозначает множество объектов целевой вредоносной активности, сохраненной в тестовой выборке, но отсутствовавшей при обучении. В исходном режиме доля объектов этого множества, вызвавших сигнал, определяется выражением:

$$\rho_{ML} = \frac{1}{|\Omega_{\text{цел}}|} \sum_{j \in \Omega_{\text{цел}}} I(\hat{y}_j = 1). \quad (15)$$

После включения интерпретационного контура сигнал формируется как для состояния «угроза», так и для состояния «подозрительное». Поэтому охват целевой активности принимает вид:

$$\begin{aligned} \rho_{ML+XAI} &= \frac{1}{|\Omega_{\text{цел}}|} \sum_{j \in \Omega_{\text{цел}}} I(Rep_j \in \\ &\in \{\text{угроза, подозрительное}\}) = \\ &= \frac{1}{|\Omega_{\text{цел}}|} \sum_{j \in \Omega_{\text{цел}}} ALERT_j. \end{aligned} \quad (16)$$

Различие между ρ_{ML} и ρ_{ML+XAI} характеризует не число подтвержденных атак, а изменение доли целевой активности, сохранившейся в области наблюдения. Расширение этой области сопровождается дополнительной проверкой объектов, которым базовая модель присвоила прогноз нормы:

$$\begin{aligned} \rho_{\text{доп}} &= \frac{1}{N_{\text{test}}} \sum_{j=1}^{N_{\text{test}}} I(\hat{y}_j = \\ &= 0 \wedge Rep_j = \text{подозрительное}). \end{aligned} \quad (17)$$

Влияние такой проверки на штатные данные определяется отдельно:

$$\rho_{\text{норм}} = \frac{\sum_{j=1}^{N_{\text{test}}} I(y_j = 0 \wedge Rep_j = \text{подозрительное})}{\sum_{j=1}^{N_{\text{test}}} I(y_j = 0)}. \quad (18)$$

Здесь N_{test} обозначает объем тестовой выборки, $\hat{y}$ — прогноз базовой ML-модели, а y_j фактическая метка. Полученные значения приведены в табл. 3.

Сопоставление двух режимов показало, что интерпретационный контур изменяет наблюдаемость целевой активности. Часть объектов, не вызвавших сигнал базовой модели, после оценки свойств объяснения получила подозрительное состояние и была сохранена для дальнейшего анализа. Этот эффект наблюдался во всех сценариях, хотя его выраженность зависела от ХАИ-конфигурации. Вместе с целевой активностью на дополнительную проверку направлялась и часть штатных объектов. Таким образом, промежуточное состояние расширяет область наблюдения, но не подтверждает наличие атаки. Выявленное соотношение между охватом целевой активности и объемом проверки необходимо учитывать при разработке метода формирования диагностического профиля и настройки порогов.

Отдельно проверялось предположение о том, что совместное использование ХАИ-методов способно повысить диагностическую информативность по сравнению с отдельным объяснителем. Для этого сопоставлялись одиночные методы, попарные конфигурации и общий ансамбль, объединяющий все пять объяснителей. Результаты представлены в табл. 4.

Попарное объединение атрибуций не обеспечило одинакового преимущества во всех сценариях. Для DDoS и PortScan лучшие попарные конфигурации несколько превосходили отдельные ХАИ-методы, тогда как в совместном сценарии наиболее информативным оказался Feature Ablation. Ансамбль из пяти объяснителей также не показал устойчивого преимущества. Полученный результат означает, что диагностическая информативность совместной конфигурации определяется взаимным дополнением объяснений, а не числом объединяемых методов. Поэтому частная гипотеза о преимуществе совместного использования ХАИ-методов получила частичное подтверждение.

Результативность совместного использования XAI-методов

Целевой сценарий	AUC одиночного XAI	AUC парной конфигурации	AUC ансамбля из пяти XAI
DDoS	0,769—0,923	0,833—0,932	0,900
PortScan	0,586—0,708	0,543—0,710	0,568
DoS GoldenEye + FTP-Patator + SSH-Patator	0,580—0,800	0,532—0,794	0,594

Проведенное исследование подтвердило, что локальное объяснение обладает измеримыми свойствами, которые не представлены в выходном классе ML-модели. Включение показателей верности и устойчивости в диагностическое описание позволило сохранить для дополнительной проверки часть целевой активности, не вызвавшей сигнал базовой модели. При этом подозрительное состояние не использовалось как фактическая метка угрозы. Результаты также показали, что информативность совместного использования XAI-методов определяется взаимным дополнением объяснений и не возрастает автоматически при увеличении числа объяснителей.

Заключение

В работе разработана интерпретационная модель наблюдения и диагностирования состояния элементов блокчейн-систем на основе применения моделей машинного обучения в условиях появления угроз нового типа. Основное отличие модели состоит во включении результата локальной интерпретации в формирование диагностического состояния отдельного объекта. Свойства объяснения получают количественное представление, после чего используются совместно с решением классификатора. Исходная ML-модель при этом не изменяется, а решения с нетипичным диагностическим профилем сохраняются для дополнительной проверки.

Экспериментальная проверка показала, что метрики верности и устойчивости отражают разные свойства локального объяснения. Их совместное рассмотрение позволяет различать характерные и нетипичные сочетания прогноза и объяснения. При поступлении вредоносной активности, признаки которой не были представлены в обучающих данных, часть объектов, отнесенных базовой моделью к норме, переводилась в подозрительное состояние. Вместе с ними на проверку направлялась часть объектов нормы. Поэтому результат интерпретационного контроля не заменяет заклю-

чение аналитика, а выделяет решения, для которых имеющихся оснований недостаточно для продолжения полностью автоматической обработки.

Проведенное исследование позволило установить три основных результата. Во-первых, измеримые свойства локального объяснения образуют дополнительное диагностическое описание решения ML-модели, которое не сводится к предсказанному классу. Во-вторых, промежуточное диагностическое состояние позволяет сохранить наблюдаемость объектов при несоответствии прогноза характеру его объяснения, не вводя новый класс атаки и не изменяя базовую модель. В-третьих, информативность совместного использования XAI-методов определяется взаимным дополнением объяснений, а не числом объединяемых методов. Парные конфигурации и ансамбль из пяти объяснителей в отдельных случаях превосходили одиночные методы, но не обеспечили устойчивого преимущества во всех сценариях.

Научный результат работы состоит в формализации связи между решением ML-модели и результатом его локальной интерпретации в задаче наблюдения и диагностирования состояния элемента блокчейн-системы. Предложенная модель определяет переход от объяснения к его измеримому описанию, а затем к диагностическому состоянию, сигналу и объяснимому отчету. Благодаря этому сведения XAI участвуют в формировании результата наблюдения, не подменяя прогноз классификатора и заключение аналитика.

В дальнейшем предполагается разработать метод интерпретационного диагностирования состояния элементов блокчейн-систем. Метод должен представить установленные моделью отношения в виде последовательности действий по формированию диагностического профиля, настройке порогов и определению состояния наблюдаемого элемента. При его экспериментальном исследовании необходимо оценить влияние выбранных правил на охват угроз нового типа и объем дополнительной проверки.

Результаты получены при финансовой поддержке проекта «Технологии противодействия ранее неизвестным квантовым киберугрозам», реализуемого в рамках государственной программы федеральной территории «Сириус» «Научно-технологическое развитие федеральной территории «Сириус» (Соглашение от 27.09.2024 г. № 23-03).

Литература

1. Ларина М.В., Скиба В.Ю. Модель цифровой платформы для анализа устойчивости к квантовым угрозам // Правовая информатика. 2025. № 3. С. 12—26.
2. Skiba V.Yu. et al. Concept of ensuring the resilience of operation of national digital platforms and blockchain ecosystems under the new quantum threat to security // Computing, Telecommunication and Control. 2025. T. 18, № 2. С. 56—73.
3. Петренко С.А., Балябин А.А. Модель квантовых угроз безопасности информации для национальных блокчейн-экосистем и платформ // Вопросы кибербезопасности. 2025. № 1 (65). С. 7—17.
4. Kayikci S., Khoshgoftaar T.M. Blockchain meets machine learning: a survey // Journal of Big Data. 2024. T. 11, № 1. Article 9.
5. Hasan M. et al. Detecting anomalies in blockchain transactions using machine learning classifiers and explainability analysis // Blockchain: Research and Applications. 2024. T. 5, № 3. Article 100207.
6. Петренко С.А., Сундеев П.В. Контроль требований к защите информации на основе модели логического вывода и графовой нейронной сети // Правовая информатика. 2026. № 2. С. 85—94.
7. Бирюков Д.Н., Дудкин А.С. Объяснимость и интерпретируемость — важные аспекты безопасности решений, принимаемых интеллектуальными системами (обзорная статья) // Научно-технический вестник информационных технологий, механики и оптики. 2025. T. 25, № 3. С. 373—386.
8. Gnidko K., Lisov D. Public Dataset and Explainable AI for Cognitive Cybersecurity: Grad-CAM Insights into Negative Visual Stimuli Detection // 2025 IEEE Ural-Siberian Conference on Biomedical Engineering, Radioelectronics and Information Technology (USBREIT). 2025. С. 17—20.
9. Hosain Y., Çakmak M. XAI-XGBoost: an innovative explainable intrusion detection approach for securing internet of medical things systems // Scientific reports. Nature Publishing Group UK London, 2025. T. 15, № 1. Article 22278.
10. Krishnan D., Singh S., Sugumaran V. Explainable AI for zero-day attack detection in IoT networks using attention fusion model // Discover Internet of Things. 2025. T. 5, № 1. Article 83.
11. Kaya K. et al. X-cba: Explainability aided catboosted anomal-e for intrusion detection system // ICC 2024-IEEE International Conference on Communications. IEEE, 2024. С. 2288—2293.
12. Senevirathna T. et al. From Insight to Action: XAI-Enhanced Detection of DDoS Attacks in Software Defined Networks // 2025 IEEE 36th International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC). IEEE, 2025. С. 1—6.
13. Villegas-Ch W. et al. Autonomous federated defense for zero-day threats in IIoT: explainable agents with real-time edge inference // Frontiers in Communications and Networks. 2026. T. 6. Article 1697204.
14. Mougan C. et al. Explanation Shift: How Did the Distribution Shift Impact the Model? // Transactions on Machine Learning Research. 2025.
15. Decker T. et al. Explanatory Model Monitoring to Understand the Effects of Feature Shifts on Performance // Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2024, Barcelona, Spain, August 25—29, 2024 / под ред. Baeza-Yates R., Bonchi F. ACM, 2024. С. 550—561.
16. Nauta M. et al. From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI // ACM Comput. Surv. 2023. T. 55, № 13s. С. 1—42.
17. Hedström A. et al. Quantus: An Explainable AI Toolkit for Responsible Evaluation of Neural Network Explanations and Beyond // Journal of Machine Learning Research. 2023. T. 24, № 34. С. 1—11.
18. Weber R.O. et al. XAI is in trouble // AI Magazine. 2024. T. 45, № 3. С. 300—316.
19. Sharafaldin I., Lashkari A.H., Ghorbani A.A. Toward generating a new intrusion detection dataset and intrusion traffic characterization // ICISp. 2018. T. 1, № 2018. С. 108—116.

AN INTERPRETATIVE MODEL FOR MONITORING AND DIAGNOSING THE STATUS OF BLOCKCHAIN SYSTEM COMPONENTS USING MACHINE LEARNING MODELS

Lisov D.N.⁵

Keywords: explainable artificial intelligence, local feature attributions, binary threat classification, explanation quality, diagnostic profile, prediction confidence, suspicious activity, analytical verification, new security threats, network intrusions.

Abstract

Purpose of work: to develop an interpretative model for monitoring and diagnosing the status of blockchain system components when new types of threats emerge, by jointly analysing the predictions of machine learning models and the results of their local interpretation.

Research method: functional decomposition was performed for the monitoring process, which includes obtaining a prediction, constructing a local explanation, quantitatively assessing its properties, and forming a diagnostic conclusion. Two quality metrics, faithfulness and robustness, were used to quantitatively assess the local explanation. The faithfulness metric shows how much replacing the features identified by the explainer changes the model output compared with replacing a random set of features, while the robustness metric characterizes the preservation of the set of important features in nearby objects with comparable predictions. Metric values were determined by the controlled replacement of input features and comparison of local explanations, and their diagnostic informativeness was assessed using receiver operating characteristic curves. To reproduce a new type of threat experimentally, one attack type was excluded from the training data and retained in the test set, after which individual explanation methods and their combinations were evaluated.

Results of the study: an interpretative model is proposed in which the result of a local explanation is incorporated into the observation of the ML model's decision and the determination of the object's diagnostic status. The measurable properties of the explanation are transformed into a vector of indicators, which is used to determine the confidence level. If the diagnostic profile of the explanation does not correspond to the characteristic profiles of known correct decisions, the object is assigned an intermediate 'suspicious' status and retained for further verification. An experimental study has shown that the explanation quality metrics contain diagnostic information not present in the model's output class; however, the extent of this information depends on the data and the explainer used. When an activity not present in the training data was encountered, the interpretation loop made it possible to retain within the observation domain some of the objects that had been missed by the base model. The combined use of several explanation methods proved useful only when the resulting attributions complemented one another and did not provide a universal advantage. The results obtained confirm that it is possible to identify decisions requiring further verification when previously unobserved threats arise. The base classifier remains unchanged, and a suspicious state is not equated with an attack.

References

1. Larina M.V., Skiba V.Iu. Model' tsifrovoy platformy dlia analiza ustoychivosti k kvantovym ugrozam // Pravovaya informatika. 2025. No. 3. S. 12—26.
2. Skiba V.Yu. et al. Concept of ensuring the resilience of operation of national digital platforms and blockchain ecosystems under the new quantum threat to security // Computing, Telecommunication and Control. 2025. T. 18. No. 2. S. 56—73.
3. Petrenko S.A., Baliabin A.A. Model' kvantovykh ugroz bezopasnosti informatsii dlia natsional'nykh blokchein-ekosistem i platform // Voprosy kiberbezopasnosti. 2025. No. 1 (65). S. 7—17.
4. Kayikci S., Khoshgoftaar T.M. Blockchain meets machine learning: a survey // Journal of Big Data. 2024. T. 11. No. 1. Article 9.

⁵ **Dmitry N. Lisov**, Ph.D. student, Junior Researcher at the Scientific Center for Information Technology and Artificial Intelligence, Sirius University of Science and Technology, Sirius, Krasnodar region, Russian Federation. ORCID: 0000-0002-8214-4109. E-mail: lisov.dn@talantiuspeh.ru

5. Hasan M. et al. Detecting anomalies in blockchain transactions using machine learning classifiers and explainability analysis // Blockchain: Research and Applications. 2024. T. 5. No. 3. Article 100207.
6. Petrenko S.A., Sundeev P.V. Kontrol' trebovaniy k zashchite informatsii na osnove modeli logicheskogo vyvoda i grafovoi neironnoi seti // Pravovaia informatika. 2026. No. 2. S. 85—94.
7. Biriukov D.N., Dudkin A.S. Ob"iasnimost' i interpretiruemost' — vazhnye aspekty bezopasnosti reshenii, prinimaemykh intellektual'nymi sistemami (obzornaia stat'ia) // Nauchno-tekhnicheskii vestnik informatsionnykh tekhnologii, mekhaniki i optiki. 2025. T. 25. No. 3. S. 373—386.
8. Gnidko K., Lisov D. Public Dataset and Explainable AI for Cognitive Cybersecurity: Grad-CAM Insights into Negative Visual Stimuli Detection // 2025 IEEE Ural-Siberian Conference on Biomedical Engineering, Radioelectronics and Information Technology (USBREIT). 2025. S. 17—20.
9. Hosain Y., Çakmak M. XAI-XGBoost: an innovative explainable intrusion detection approach for securing internet of medical things systems // Scientific reports. Nature Publishing Group UK London, 2025. T. 15. No. 1. Article 22278.
10. Krishnan D., Singh S., Sugumaran V. Explainable AI for zero-day attack detection in IoT networks using attention fusion model // Discover Internet of Things. 2025. T. 5. No. 1. Article 83.
11. Kaya K. et al. X-cba: Explainability aided catboosted anomal-e for intrusion detection system // ICC 2024-IEEE International Conference on Communications. IEEE, 2024. S. 2288—2293.
12. Senevirathna T. et al. From Insight to Action: XAI-Enhanced Detection of DDoS Attacks in Software Defined Networks // 2025 IEEE 36th International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC). IEEE, 2025. S. 1—6.
13. Villegas-Ch W. et al. Autonomous federated defense for zero-day threats in IIoT: explainable agents with real-time edge inference // Frontiers in Communications and Networks. 2026. T. 6. Article 1697204.
14. Mougan C. et al. Explanation Shift: How Did the Distribution Shift Impact the Model? // Transactions on Machine Learning Research. 2025.
15. Decker T. et al. Explanatory Model Monitoring to Understand the Effects of Feature Shifts on Performance // Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2024, Barcelona, Spain, August 25—29, 2024 / pod red. Baeza-Yates R., Bonchi F. ACM, 2024. S. 550—561.
16. Nauta M. et al. From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI // ACM Comput. Surv. 2023. T. 55. No. 13s. S. 1—42.
17. Hedström A. et al. Quantus: An Explainable AI Toolkit for Responsible Evaluation of Neural Network Explanations and Beyond // Journal of Machine Learning Research. 2023. T. 24. No. 34. S. 1—11.
18. Weber R.O. et al. XAI is in trouble // AI Magazine. 2024. T. 45. No. 3. S. 300—316.
19. Sharafaldin I., Lashkari A.H., Ghorbani A.A. Toward generating a new intrusion detection dataset and intrusion traffic characterization // ICISp. 2018. T. 1. No. 2018. S. 108—116.

