

МЕТОД ИНТЕРПРЕТАЦИОННОГО ДИАГНОСТИРОВАНИЯ СОСТОЯНИЯ ЭЛЕМЕНТОВ БЛОКЧЕЙН-СИСТЕМ ПО РЕШЕНИЯМ МОДЕЛЕЙ МАШИННОГО ОБУЧЕНИЯ

Лисов Д.Н.¹, Гнидко К.О.²

Ключевые слова: объяснимый искусственный интеллект, оценка качества объяснения, локальное объяснение, статистическое обучение, диагностические правила, диагностический профиль, калибровка порогов, подозрительное состояние, угрозы нового типа.

Аннотация

Цель работы: разработать метод интерпретационного диагностирования состояния элементов блокчейн-систем по решениям моделей машинного обучения при появлении угроз нового типа.

Методы исследования: различия между правильными и ошибочными решениями ML-модели исследовались по показателям локального объяснения. Эти показатели характеризовали изменение выхода модели при контролируемой замене входных признаков, структуру атрибуций и согласованность результатов нескольких объяснителей. Полученные диагностические профили использовались для построения правил методами статистического обучения. Пороговые границы определялись в соответствии со статистической теорией принятия решений при ограничении доли корректных прогнозов, направляемых на дополнительную проверку. Различимость правильных и ошибочных решений оценивалась по AUC на независимых данных. Практический результат характеризовался изменением охвата ранее не наблюдавшейся вредоносной активности, концентрацией пропущенных угроз в подозрительной области и объёмом дополнительной проверки.

Результаты исследования: разработан метод, дополняющий бинарный прогноз ML-модели диагностическим состоянием, формируемым по характеристикам локального объяснения без изменения исходного классификатора. Научная новизна состоит в построении отдельных диагностических правил для положительной и отрицательной ветвей прогноза и в независимой калибровке их порогов с учётом допустимой доли корректных решений, направляемых на дополнительную проверку. Экспериментальное исследование на сетевых данных и различных представлениях смарт-контрактов показало, что сформированные правила различают правильные и ошибочные решения при разных XAI-конфигурациях. В сценариях появления вредоносной активности, не использованной при обучении ML-модели и построении диагностических правил, часть пропущенных объектов была переведена в подозрительное состояние. Одновременно определялся возникающий объём дополнительной проверки. Результаты подтвердили диагностическую информативность показателей локального объяснения и показали её зависимость от данных, ML-модели и XAI-конфигурации. Универсально предпочтительная XAI-конфигурация не установлена.

DOI: 10.24412/1994-1404-2026-3-29-40

Введение и постановка задачи

В исследованиях цифровых и облачных платформ выявление деструктивных воздействий и учёт новых типов угроз рассматриваются как условия устойчивого функционирования защищаемых систем [1, 2]. При включении ML-моделей в процедуры защиты возникает проблема объяснимости получае-

мых решений, поскольку достоверность автоматизированного контроля зависит не только от предсказанного класса, но и от возможности установить признаки, повлиявшие на результат [3]. Для решения этой проблемы вместе с моделями машинного обучения применяются средства объяснимого искусственного интеллекта.

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

² Гнидко Константин Олегович, доктор технических наук, доцент, заместитель руководителя группы Научного центра информационных технологий и искусственного интеллекта АНОО ВО «Научно-технологический университет «Сириус»», пгт Сириус, Краснодарский край, Российская Федерация. ORCID: 0000-0002-8605-8865.

E-mail: gnidko.ko@talantiuspeh.ru

При анализе транзакций блокчейна они позволяют определить признаки, обусловившие отнесение операции к аномальным [4]. В системах обнаружения сетевых атак локальные объяснения используются для анализа решений классификаторов, в том числе при исследовании активности, отличающейся от представленной в обучающих данных [5—7]. В этих работах ML-модель устанавливает класс объекта, после чего ХАИ раскрывает влияние входных признаков на уже полученный прогноз. Такой подход повышает прозрачность классификации, однако объяснение не участвует в определении состояния объекта.

Более тесное включение ХАИ в защитный контур предложено в работах [8, 9]. В первом случае значения атрибуций передаются дополнительному классификатору, уточняющему результат обнаружения распределённой атаки. Во втором объяснения используются для оценки поведения участников федеративного обучения и ограничения влияния потенциально скомпрометированных узлов. Другой подход основан на наблюдении за изменением распределений объяснений [10, 11]. Он позволяет связать сдвиг входных данных с изменением поведения модели, но характеризует группу поступивших объектов, а не отдельное решение. Таким образом, результаты ХАИ могут использоваться не только для последующего пояснения прогноза. Вместе с тем известные решения либо изменяют исходную схему классификации, либо решают задачу наблюдения на другом уровне и не формируют диагностическое состояние отдельного объекта по характеристикам его локального объяснения.

Возможность такого применения ХАИ связана с развитием методов количественной оценки объяснений [12—14]. Предложенные показатели характеризуют реакцию модели на изменение выделенных признаков, структуру атрибуций и согласованность результатов разных объяснителей. Каждый из них отражает отдельное свойство объяснения, а его информативность зависит от данных, ML-модели и применяемого ХАИ-метода. Поэтому значение одного показателя не может непосредственно рассматриваться как диагностическое состояние. Для совместного использования нескольких характеристик необходимо установить, как их сочетания проявляются в правильных и ошибочных решениях конкретной ML-модели.

В ранее выполненном исследовании [15] разработана интерпретационная модель наблюдения и диагностирования состояния элементов блокчейн-систем. Предусмотренная моделью последовательность связывает прогноз и локальное объяснение с диагностическим профилем, оценкой соответствия и состоянием объекта:

$$\begin{aligned} (\hat{y}_{(B_i)}(t), \text{Expl}_{(B_i)}(t)) &\rightarrow \mathcal{M}_{(B_i)}(t) \rightarrow \\ &\rightarrow \tau_{(B_i)}(t) \rightarrow (\text{Rep}_{(B_i)}(t), \text{ALERT}_{(B_i)}(t)). \end{aligned} \quad (1)$$

В выражении (1) определены состав результатов и порядок их получения, но не установлено правило g ,

преобразующее диагностический профиль $\mathcal{M}_{(B_i)}(t)$ в оценку $\tau_{(B_i)}(t)$. Не определены также пороговые границы, по которым эта оценка связывается с диагностическим состоянием. Назначить их одинаковыми для разных защищаемых элементов нельзя, поскольку распределения показателей зависят от используемых данных, поведения ML-модели и выбранной ХАИ-конфигурации. Для практической реализации модели требуется способ воспроизводимого формирования правила и его порогов применительно к зафиксированному сочетанию этих компонентов.

Исследование проводится применительно к защищаемому элементу B_i , на котором используются обученная бинарная ML-модель, зафиксированная предобработка данных и совместимая ХАИ-конфигурация. Для эталонных объектов известны фактические метки и прогнозы модели, поэтому среди них могут быть выделены правильные и ошибочные решения. При поступлении нового объекта доступны его признаки, прогноз ML-модели и локальное объяснение, тогда как фактическое состояние устанавливается позднее. Следовательно, зависимость между характеристиками объяснения и типом решения должна быть выявлена по эталонным данным, после чего полученное правило может применяться к поступающим объектам.

В основу исследования положена гипотеза о том, что для фиксированной ML-модели и выбранной ХАИ-конфигурации совместное распределение показателей локального объяснения различается у правильных и ошибочных решений. Если это различие сохраняется на независимых данных, диагностический профиль может использоваться для выделения решений, требующих дополнительной проверки, в том числе при появлении вредоносной активности, не представленной при обучении модели.

Цель работы состоит в разработке и экспериментальном исследовании метода интерпретационного диагностирования состояния элементов блокчейн-систем по решениям моделей машинного обучения. Метод должен обеспечить формирование диагностического профиля локального объяснения, построение правил для положительной и отрицательной ветвей прогноза и установление порогов, ограничивающих направление корректных решений на дополнительную проверку.

Принцип построения метода

Проверка сформулированной гипотезы требует сначала установить связь между диагностическим профилем локального объяснения и типом решения ML-модели, а затем использовать найденную зависимость для объектов, фактическое состояние которых в момент обработки неизвестно. Этим обусловлено разделение метода на подготовительный и рабочий этапы. До формирования профилей эталонная выборка разделяется на непересекающиеся части D_{fit} и D_{cal} . Первая используется для выявления закономерностей,

характеризующих правильные и ошибочные решения, тогда как вторая предназначена для независимого установления пороговых границ.

Для объектов D_{fit} их сопоставление позволяет выделить группы TP, FP, TN и FN, которые характеризуют правильные и ошибочные решения положительной и отрицательной ветвей прогноза. Различный смысл ошибок не позволяет описывать их одним правилом. В положительной ветви необходимо установить, насколько объяснение подтверждает прогноз угрозы, а в отрицательной требуется выявить сходство с известными пропусками вредоносной активности. Поэтому по диагностическим профилям объектов D_{fit} различает TP и FP, а правило $\mathfrak{D}_0$ различает TN и FN.

Получаемые правила связывают диагностический профиль с числовой оценкой τ , но границы диагностических состояний устанавливаются отдельно. Для этого к объектам D_{cal} применяются сформированные по D_{fit} профили и правила. В положительной ветви порог θ_{TP} разделяет состояния «угроза» и «подозрительное», в отрицательной порог θ_{FN} отделяет «важное» состояние от «подозрительного». Параметры α_{TP} и α_{TN} ограничивают долю правильных решений соответствующей ветви, которые могут быть направлены на дополнительную проверку. Среди удовлетворяющих этому ограничению порогов выбирается вариант, обеспечивающий наиболее полное выявление ошибочных решений. Результатом подготовительного этапа становятся правила расчёта диагностического профиля, правила $\mathfrak{D}_1$ и $\mathfrak{D}_0$, а также установленные для них пороги.

На рабочем этапе для поступившего объекта $X_{(B_i)}(t)$ формируются прогноз $\hat{y}_{(B_i)}(t)$ и локальное объяснение $Expl_{(B_i)}(t)$. По объяснению рассчитывается диагностический профиль $M_{(B_i)}(t)$, а прогноз определяет выбор правила $\mathfrak{D}_1$ или $\mathfrak{D}_0$. Полученная оценка $\tau_{(B_i)}(t)$ сопоставляется с порогом соответствующей ветви,

после чего устанавливается диагностическое состояние $Rep_{(B_i)}(t)$. На его основе формируются сигнал $ALERT_{(B_i)}(t)$ и объяснимый отчет $Report_{(B_i)}(t)$. Взаимосвязь подготовительного и рабочего этапов представлена на рис. 1.

Формирование диагностического профиля объяснения

Диагностические правила $\mathfrak{D}_1$ и $\mathfrak{D}_0$ должны различать правильные и ошибочные решения по свойствам локального объяснения. Исходные атрибуции для этого недостаточны, поскольку они показывают вклад признаков, но не характеризуют его соответствие реакции ML-модели и известным типам решений. Для формирования такого описания объекты D_{fit} распределяются по результатам сопоставления фактической метки y и прогноза $\hat{y}$:

$$\begin{aligned} TP &= \{X_{(B_i)} \in D_{fit}: y_{(B_i)} = 1, \hat{y}_{(B_i)} = 1\}, \\ FP &= \{X_{(B_i)} \in D_{fit}: y_{(B_i)} = 0, \hat{y}_{(B_i)} = 1\}, \\ TN &= \{X_{(B_i)} \in D_{fit}: y_{(B_i)} = 0, \hat{y}_{(B_i)} = 0\}, \\ FN &= \{X_{(B_i)} \in D_{fit}: y_{(B_i)} = 1, \hat{y}_{(B_i)} = 0\}. \end{aligned}$$

Группы TP и TN содержат правильные решения положительной и отрицательной ветвей, а FP и FN содержат ошибки этих ветвей. Локальное объяснение объекта, полученное r -м XAI-методом, представляется вектором атрибуций:

$$\begin{aligned} \varepsilon_r^{(B_i)}(t) &= in_r(X_{(B_i)}(t), \hat{y}_{(B_i)}(t)) = \\ &= (a_{1,r}^{(B_i)}(t), a_{2,r}^{(B_i)}(t), \dots, a_{k,r}^{(B_i)}(t)), \\ &r = 1, \dots, p, \end{aligned}$$

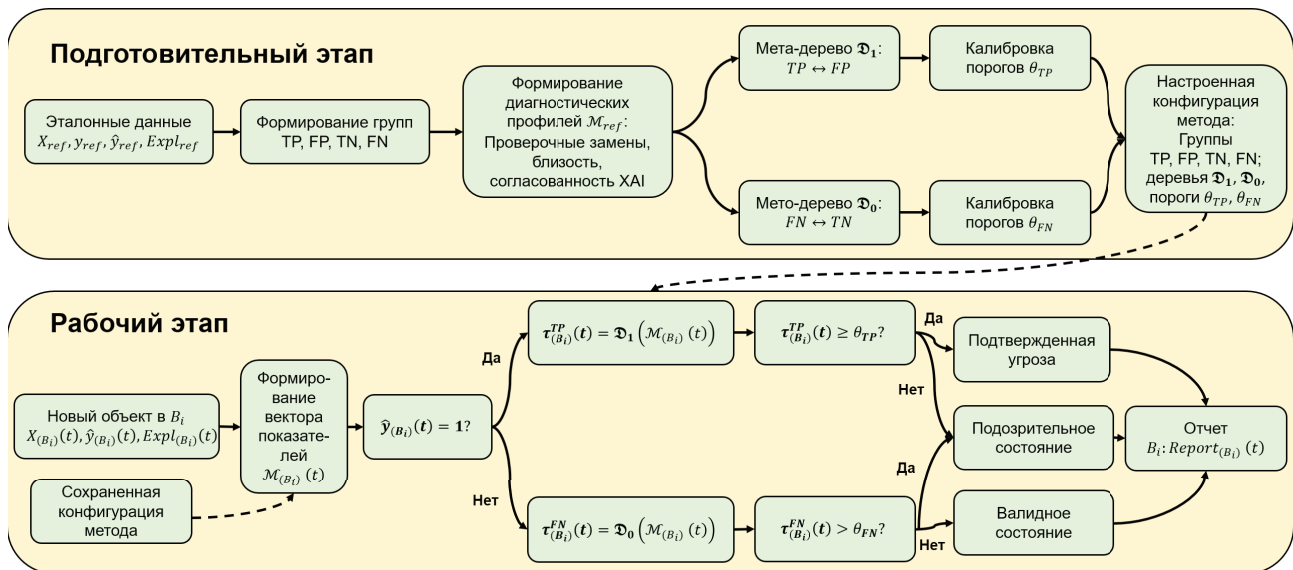


Рис. 1. Общая схема метода интерпретационного диагностирования состояния элемента блокчейн-системы

где a характеризует направление и величину вклада признака x в прогноз объекта $X_{(B_i)}(t)$, k — число входных признаков, p — число ХАИ-методов в используемой конфигурации. Соответствие атрибуций поведению ML-модели проверяется заменой выбранных признаков значениями ближайших правильных решений той же ветви прогноза. Источником замен служит группа:

$$Cand_{(B_i)}(t) = \begin{cases} TP_{(B_i)}, & \hat{y}_{(B_i)}(t) = 1, \\ TN_{(B_i)}, & \hat{y}_{(B_i)}(t) = 0. \end{cases}$$

После зафиксированной предобработки кандидаты упорядочиваются по расстоянию:

$$d(X_{(B_i)}(t), X_{(B_i),cand}) = \sqrt{\sum_{l=1}^k (x_l - x_{l,cand})^2}. \quad (2)$$

Совпадающий с исследуемым объект исключается. Первые R кандидатов образуют последовательность локальных замен:

$$X_{(B_i)}^{rep}(t) = \{X_{(B_i),cand_1}, \dots, X_{(B_i),cand_R}\}.$$

Положительные и отрицательные атрибуции отбираются отдельно, чтобы проверить влияние каждого направления на выход модели. Предельное число рассматриваемых признаков определяется выражением:

$$K = \max(k_{min}, [\rho k]),$$

где ρ — доля анализируемых признаков, k_{min} — их минимально допустимое число. Множество $P_r^{(B_i)}(t)$ содержит индексы не более K положительных атрибуций r -го объяснителя с наибольшими абсолютными значениями. Множество $N_r^{(B_i)}(t)$ формируется аналогично среди отрицательных атрибуций.

Для каждого объекта $X \in X_{(B_i)}^{rep}(t)$ и множества индексов $Index \in \{P_r^{(B_i)}(t), N_r^{(B_i)}(t)\}$ строятся два проверочных объекта. В первом заменяются признаки, выделенные объяснителем:

$$\begin{aligned} X_j^{(-Index)(B_i)} &= (x_{1,j}^{(-Index)(B_i)}, x_{2,j}^{(-Index)(B_i)}, \dots, \\ &\quad x_{k,j}^{(-Index)(B_i)}), \\ x_{l,j}^{(-Index)(B_i)}(X_{(B_i)}(t)) &= \begin{cases} x_{l,j}^{(B_i),rep}, l \in Index \\ x_l^{(B_i)}, l \notin Index \end{cases}, \end{aligned} \quad (3)$$

Во втором сохраняются выбранные признаки и заменяются остальные:

$$\begin{aligned} X_j^{(+Index)(B_i)} &= (x_{1,j}^{(+Index)(B_i)}, x_{2,j}^{(+Index)(B_i)}, \dots, \\ &\quad x_{k,j}^{(+Index)(B_i)}), \\ x_{l,j}^{(+Index)(B_i)}(X_{(B_i)}(t)) &= \begin{cases} x_{l,j}^{(B_i),rep}, l \notin Index \\ x_l^{(B_i)}, l \in Index \end{cases}. \end{aligned} \quad (4)$$

Первое изменение позволяет проверить, проявляется ли указанное объяснителем влияние в выходе модели. Второе характеризует способность выбранных признаков воспроизвести исходный результат [12, 13]. Если $z_{B_i}(X_{(B_i)}(t))$ обозначает вероятность положительного класса, реакция модели на замену признаков определяется показателями:

$$E_r^{(+)(B_i)}(X_{(B_i)}(t)) = \frac{1}{R} \sum_{j=1}^R (z_{B_i}(X_{(B_i)}(t)) - z_{B_i}(X_j^{(-P)(B_i)})), \quad (5)$$

$$E_r^{(-)(B_i)}(X_{(B_i)}(t)) = \frac{1}{R} \sum_{j=1}^R (z_{B_i}(X_j^{(-N)(B_i)}) - z_{B_i}(X_{(B_i)}(t))), \quad (6)$$

Положительное значение $E_r^{(\pm)}$ соответствуют реакции модели, согласованной с направлением выбранных атрибуций. Отклонение выхода при сохранении только этих признаков оценивается показателями достаточности:

$$S_r^{(+)(B_i)}(X_{(B_i)}(t)) = \frac{1}{R} \sum_{j=1}^R |z_{B_i}(X_{(B_i)}(t)) - z_{B_i}(X_j^{(+P)(B_i)})|, \quad (7)$$

$$S_r^{(-)(B_i)}(X_{(B_i)}(t)) = \frac{1}{R} \sum_{j=1}^R |z_{B_i}(X_{(B_i)}(t)) - z_{B_i}(X_j^{(+N)(B_i)})|, \quad (8)$$

Меньшее значение $S_r^{(\pm)}$ или $S_r^{(-)}$ означает более полное воспроизведение исходного выхода. Для совместной ХАИ-конфигурации проверочные показатели усредняются:

$$E^{(\pm)(B_i)}(X_{(B_i)}(t)) = \frac{1}{p} \sum_{r=1}^p E_r^{(\pm)(B_i)}(X_{(B_i)}(t)), \quad (9)$$

$$S^{(\pm)(B_i)}(X_{(B_i)}(t)) = \frac{1}{p} \sum_{r=1}^p S_r^{(\pm)(B_i)}(X_{(B_i)}(t)). \quad (10)$$

Различие в достаточности двух направлений выражается показателем:

$$\begin{aligned} \Delta S(X_{(B_i)}(t)) &= \\ &= S^{(-)(B_i)}(X_{(B_i)}(t)) - S^{(+)(B_i)}(X_{(B_i)}(t)), \end{aligned} \quad (11)$$

Положительное значение ΔS указывает на меньший остаток для положительных атрибуций, отрицательное значение соответствует более высокой достаточности противоположного направления.

Проверочные изменения не отражают распределение влияния между входными признаками. Для его оценки модули атрибуций преобразуются в относительные веса:

$$\omega_{l,r}^{(B_i)}(t) = \frac{|a_{l,r}^{(B_i)}(t)|}{\sum_{h=1}^k |a_{h,r}^{(B_i)}(t)|}, \quad l = 1, \dots, k. \quad (12)$$

Энтропия полученного распределения рассчитывается по ненулевым весам:

$$H_r^{(B_i)}(t) = - \sum_{\substack{l=1 \\ \omega_{l,r}^{(B_i)}(t) > 0}}^k \omega_{l,r}^{(B_i)}(t) \ln \omega_{l,r}^{(B_i)}(t). \quad (13)$$

Концентрация атрибуций определяется дополнением нормированной энтропии до единицы:

$$C_r^{(B_i)}(t) = 1 - \frac{H_r^{(B_i)}(t)}{\ln k}. \quad (14)$$

Значение $C_r^{(B_i)}(t)$ близкое к единице, соответствует сосредоточению влияния на небольшом числе признаков. При отсутствии ненулевых атрибуций принимается $C_r^{(B_i)}(t) = 0$. Для совместной ХАИ-конфигурации используется среднее значение:

$$C^{(B_i)}(t) = \frac{1}{p} \sum_{r=1}^p C_r^{(B_i)}(t). \quad (15)$$

Усреднение проверочных показателей не учитывает различия между отдельными объяснителями. Их разброс сохраняется с помощью стандартных отклонений:

$$\sigma_{E^{(+)}}^{(B_i)}(X_{(B_i)}(t)) = \sqrt{\frac{1}{p} \sum_{r=1}^p \left(E_r^{(+)(B_i)} - E^{(+)(B_i)} \right)^2}, \quad (16)$$

$$\sigma_{E^{(-)}}^{(B_i)}(X_{(B_i)}(t)) = \sqrt{\frac{1}{p} \sum_{r=1}^p \left(E_r^{(-)(B_i)} - E^{(-)(B_i)} \right)^2}, \quad (17)$$

$$\sigma_{S^{(+)}}^{(B_i)}(X_{(B_i)}(t)) = \sqrt{\frac{1}{p} \sum_{r=1}^p \left(S_r^{(+)(B_i)} - S^{(+)(B_i)} \right)^2}, \quad (18)$$

$$\sigma_{S^{(-)}}^{(B_i)}(X_{(B_i)}(t)) = \sqrt{\frac{1}{p} \sum_{r=1}^p \left(S_r^{(-)(B_i)} - S^{(-)(B_i)} \right)^2}, \quad (19)$$

Близость числовых оценок не означает, что ХАИ-методы выделили одинаковые признаки. Состав положительных и отрицательных топ-множеств сопоставляется с помощью коэффициента Жаккара:

$$J^{(+)}(X_{(B_i)}(t)) = \frac{2}{p(p-1)} \sum_{r=1}^{p-1} \sum_{h=r+1}^p \frac{|P_r^{(B_i)}(t) \cap P_h^{(B_i)}(t)|}{|P_r^{(B_i)}(t) \cup P_h^{(B_i)}(t)|}, \quad (20)$$

$$J^{(-)}(X_{(B_i)}(t)) = \frac{2}{p(p-1)} \sum_{r=1}^{p-1} \sum_{h=r+1}^p \frac{|N_r^{(B_i)}(t) \cap N_h^{(B_i)}(t)|}{|N_r^{(B_i)}(t) \cup N_h^{(B_i)}(t)|}, \quad (21)$$

При пустом объединении двух множеств соответствующее отношение принимается равным единице. Коэффициенты $J^{(\pm)}$ учитывают состав топ-множеств,

но не порядок влияния всех признаков. Пусть $Ind_{l,r}$

обозначает ранг $|a_{l,r}^{(B_i)}(t)|$ в последовательности атрибуций r -го объяснителя, причем совпадающим значениям назначается средний ранг:

$$\overline{Ind}_r = \frac{1}{k} \sum_{l=1}^k Ind_{l,r}. \quad (22)$$

Согласованность рангов объяснителей r и h определяется выражением:

$$Corr_{r,h}(X_{(B_i)}(t)) = \frac{\sum_{l=1}^k (Ind_{l,r} - \overline{Ind}_r)(Ind_{l,h} - \overline{Ind}_h)}{\sqrt{\sum_{l=1}^k (Ind_{l,r} - \overline{Ind}_r)^2} \sqrt{\sum_{l=1}^k (Ind_{l,h} - \overline{Ind}_h)^2}}, \quad (23)$$

Корреляция определяется для пар ХАИ-методов, у которых обе последовательности рангов имеют ненулевой разброс. Множество таких пар обозначается $\mathcal{P}_{rank}(t)$, а их количество — $kol_{rank}(t)$. Средняя ранговая согласованность составляет:

$$A_{rank}(X_{(B_i)}(t)) = \frac{2}{kol_{rank}(t)} \sum_{(r,h) \in \mathcal{P}_{rank}(t)} Corr_{r,h}(X_{(B_i)}(t)). \quad (24)$$

Если ранговую корреляцию невозможно вычислить ни для одной пары ХАИ-методов, принимается $A_{rank} = 0$. Ранговая согласованность рассчитывается по модулям атрибуций и потому не учитывает направление влияния признаков. Для его сопоставления используется знаковая функция:

$$sign(a) = \begin{cases} 1, & a > 0, \\ 0, & a = 0, \\ -1, & a < 0. \end{cases}$$

В оценку включаются признаки, которым хотя бы один ХАИ-метод назначил ненулевую атрибуцию. Их активность обозначается $u_l^{(B_i)}$, а количество — $k_a^{(B_i)}$:

$$u_l^{(B_i)}(t) = I \left(\sum_{r=1}^p |a_{l,r}^{(B_i)}(t)| > 0 \right), \quad (25)$$

$$k_a^{(B_i)}(t) = \sum_{l=1}^k u_l^{(B_i)}(t). \quad (26)$$

Для каждого активного признака сопоставляются направления, полученные отдельными объяснителями. Их средняя согласованность определяется выражением:

$$A_{sign}(X_{(B_i)}(t)) = \frac{1}{k_a^{(B_i)}(t)} \sum_{l=1}^k u_l^{(B_i)}(t) \left| \frac{1}{p} \sum_{r=1}^p sign(a_{l,r}^{(B_i)}(t)) \right|, \quad (27)$$

Если $k_a^{(B_i)}(t) = 0$, принимается $A_{sign} = 0$. Значение показателя уменьшается при расхождении направлений и при отсутствии атрибуции у части объяснителей.

Полученные показатели характеризуют объяснение, но не положение самого объекта относительно известных решений. Для каждой группы

$G \in \{TP, FP, TN, FN\}$ определяется расстояние до ближайшего объекта:

$$d_G(X_{(B_i)}(t)) = \min_{X_{(B_i),c} \in G, X_{(B_i),c} \neq X_{(B_i)}(t)} \sqrt{\sum_{l=1}^k (x_l - x_{l,c})^2}, \quad (28)$$

Относительное положение между правильной и ошибочной группами соответствующей ветви выражается показателями:

$$Near_{FN}(X_{(B_i)}(t)) = d_{TN} - d_{FN}, \quad (29)$$

$$Near_{FP}(X_{(B_i)}(t)) = d_{TP} - d_{FP}, \quad (30)$$

Положительное значение $Near_{FN}$ и $Near_{FP}$ соответствуют большей близости объекта к известным ошибочным решениям FN и FP.

Совместное использование показателей требует привести их к сопоставимому диапазону. Преобразования определяются по их эмпирическим распределениям на объектах D_{fit} . Для исходного показателя $\tilde{m}_s$, увеличение которого используется в прямом направлении, применяется выражение:

$$m_s^{(B_i)}(t) = \frac{1}{kol_{fit}} \sum_{k_{fit}=1}^{kol_{fit}} I(\tilde{m}_{s,k_{fit}}^{(B_i)}(t) \leq \tilde{m}_s^{(B_i)}(t)). \quad (31)$$

где kol_{fit} – число объектов D_{fit} . Прямое преобразование применяется к $E^{(\pm)}$, ΔS , $Near_{FN}$ и $Near_{FP}$. Для показателей $S^{(\pm)}$ и σ используется обратное направление шкалы:

$$m_s^{(B_i)}(t) = 1 - \frac{1}{kol_{fit}} \sum_{k_{fit}=1}^{kol_{fit}} I(\tilde{m}_{s,k_{fit}}^{(B_i)}(t) \leq \tilde{m}_s^{(B_i)}(t)). \quad (32)$$

Показатели C , $J^{(\pm)}$ и A_{sign} уже принадлежат диапазону $[0,1]$. Значение A_{rank} линейно переносится из диапазона $[-1,1]$ в $[0,1]$. Распределения D_{fit} используются при преобразовании показателей калибровочных, тестовых и поступающих объектов.

При использовании одного XAI-метода диагностический профиль имеет вид:

$$\mathcal{M}_{(B_i)}(t) = (E^{(\pm)}, S^{(\pm)}, \Delta S, C, Near_{FN}, Near_{FP}).$$

При совместной XAI-конфигурации он дополняется компонентами: $\sigma_{E^{(\pm)}}$, $\sigma_{S^{(\pm)}}$, $J^{(\pm)}$, A_{rank} , A_{sign} .

Сформированный профиль характеризует реакцию ML-модели на проверочные изменения, распределение влияния между признаками, согласованность объяснителей и положение объекта относительно известных типов решений. Совместное проявление этих свойств образует признаковое описание для синтеза правил $\mathfrak{D}_1$ и $\mathfrak{D}_0$.

Синтез диагностического правила

Различие правильных и ошибочных решений может определяться сочетанием нескольких компонентов диагностического профиля, поэтому оно не сводится к пороговой проверке отдельного показателя. Искомая зависимость устанавливается в пространстве сформированных профилей. Прогноз ML-модели разделяет эту задачу на две ветви. При $\hat{y}_{(B_i)}(t) = 1$ сопоставляются решения TP и FP, а при $\hat{y}_{(B_i)}(t) = 0$ сопоставляются TN и FN. Различный диагностический смысл этих сравнений приводит к построению правил $\mathfrak{D}_1$ и $\mathfrak{D}_0$.

Обучающее множество положительной ветви включает профили объектов TP и FN:

$$\mathcal{M}_{(B_i)}^{fit-\mathfrak{D}_1} = \{(\mathcal{M}_{(B_i),k_{fit}}(t), Class_{k_{fit}}^{\mathfrak{D}_1}) : TP \cup FP\},$$

где

$$Class_{k_{fit}}^{\mathfrak{D}_1} = \begin{cases} 1, & X_{(B_i)}^{k_{fit}} \in TP \\ 0, & X_{(B_i)}^{k_{fit}} \in FP \end{cases}.$$

Положительный класс представлен правильно выявленными угрозами. Поэтому выход $\mathfrak{D}_1$ характеризует соответствие исследуемого профиля сочетаниям показателей, наблюдавшимся у TP. Снижение этой оценки указывает на смещение профиля в сторону ложноположительных решений.

Для отрицательной ветви формируется множество:

$$\mathcal{M}_{(B_i)}^{fit-\mathfrak{D}_0} = \{(\mathcal{M}_{(B_i),k_{fit}}(t), Class_{k_{fit}}^{\mathfrak{D}_0}) : TN \cup FN\},$$

где

$$Class_{k_{fit}}^{\mathfrak{D}_0} = \begin{cases} 0, & X_{(B_i)}^{k_{fit}} \in TN \\ 1, & X_{(B_i)}^{k_{fit}} \in FN \end{cases}.$$

В этой ветви положительный класс соответствует пропускам угроз. Выход $\mathfrak{D}_0$ характеризует сходство исследуемого профиля с сочетаниями показателей, наблюдавшимися у FN. Формирование каждого правила требует наличия обеих сравниваемых групп в D_{fit} .

Различия между группами могут одновременно проявляться по нескольким координатам профиля. Для представления таких сочетаний используются деревья решений (Decision Trees), последовательно разделяющие пространство $\mathcal{M}_{(B_i)}$. Неравное число правильных и ошибочных решений компенсируется весами:

$$\varphi_{TP} = \frac{1}{2Kol_{fit,TP}}, \quad \varphi_{FP} = \frac{1}{2Kol_{fit,FP}},$$

$$\varphi_{TN} = \frac{1}{2Kol_{fit,TN}}, \quad \varphi_{FN} = \frac{1}{2Kol_{fit,FN}},$$

где $Kol_{fit,G}$ обозначает число объектов группы G в D_{fit} . При таком распределении суммарный вес каждой из двух групп соответствующей ветви равен $1/2$. Правила формируются выражениями:

$$\mathfrak{D}_1 = \text{DecisionTreeClassifier} \left(\mathcal{M}_{(B_i)}^{fit_{\mathfrak{D}_1}}, \varphi_{TP}, \varphi_{FP} \right), \quad (33)$$

$$\mathfrak{D}_0 = \text{DecisionTreeClassifier} \left(\mathcal{M}_{(B_i)}^{fit_{\mathfrak{D}_0}}, \varphi_{TN}, \varphi_{FN} \right). \quad (34)$$

Глубина дерева и минимальное число объектов в листе ограничивают сложность получаемых правил. Внутренние условия выбираются по координатам диагностического профиля, а каждому листу ставится в соответствие оценка, определяемая взвешенным соотношением сравниваемых классов.

Полученные оценки еще не задают границы диагностических состояний. Для их установления применяются объекты D_{cal} , профили которых рассчитываются с использованием эталонных групп и эмпирических преобразований, сформированных по D_{fit} . В положительной и отрицательной ветвях вычисляются оценки:

$$\tau_{(B_i),TP,k_{cal}}^{cal_{\mathfrak{D}_1}} = \mathfrak{D}_1 \left(\mathcal{M}_{(B_i),k_{cal}}^{cal_{\mathfrak{D}_1}} \right), \quad (35)$$

$$\tau_{(B_i),FN,k_{cal}}^{cal_{\mathfrak{D}_0}} = \mathfrak{D}_0 \left(\mathcal{M}_{(B_i),k_{cal}}^{cal_{\mathfrak{D}_0}} \right). \quad (36)$$

Обе оценки принадлежат диапазону $[0,1]$, но имеют разную смысловую направленность. Увеличение $\tau_{(B_i),TP}^{cal_{\mathfrak{D}_1}}$

означает большее соответствие профилю TP, тогда как увеличение $\tau_{(B_i),FN}^{cal_{\mathfrak{D}_0}}$ означает большее соответствие

профилю FN. Поэтому в положительной ветви подозрительная область задается условием $\tau_{(B_i),TP}^{cal_{\mathfrak{D}_1}} < \theta_{TP}$, а в отрицательной условием $\tau_{(B_i),FN}^{cal_{\mathfrak{D}_0}} > \theta_{FN}$. Оценки дерева

принимают конечное число значений, поэтому все возможные варианты разделения калибровочных объектов могут быть представлены конечной последовательностью порогов. Пусть различные оценки соответствующей ветви упорядочены как $\tau_1 < \tau_2 < \dots < \tau_{kol_{cal}}$. В последовательность кандидатов включаются границы диапазона 0 и 1, значения τ , а также середины интервалов между соседними оценками $\frac{\tau_{k_{cal}} + \tau_{k_{cal}+1}}{2}$. Повторяющиеся значения исключаются, после чего кандидаты упорядочиваются по возрастанию. Кандидат обозначается θ_j^{cand} , а их общее число — kol_{cand} .

Каждый кандидат определяет состав подозрительной области, в которую вместе с ошибочными могут быть направлены правильные решения ML-модели. Поэтому допустимость порога сначала оценивается по доле правильных решений соответствующей ветви, отнесенных к подозрительному состоянию. Для положительной ветви эта доля составляет:

Показатели FPR_{TP} и FPR_{TN} выражают доли правильных решений соответствующих ветвей, направленных в подозрительное состояние. Допустимыми признаются кандидаты, для которых выполняются условия:

$$FPR_{TN}^{\theta_j^{cand}} = \frac{1}{kol_{cal,TN}} \sum_{k_{cal}: X_{k_{cal}}^{cal} \in \in TN} I \left(\tau_{(B_i),FN,k_{cal}}^{cal_{\mathfrak{D}_0}} > \theta_j^{cand} \right). \quad (38)$$

$$FPR_{TP}^{\theta_j^{cand}} \leq \alpha_{TP}, \quad FPR_{TN}^{\theta_j^{cand}} \leq \alpha_{TN}, \quad (39)$$

где α_{TP} и α_{TN} задают предельные доли правильных решений соответствующих ветвей, допускаемые к дополнительной проверке. Среди удовлетворяющих этим ограничениям кандидатов порог выбирается по полноте выделения ошибочной группы. Для положительной ветви полнота обнаружения FP определяется выражением:

$$Recall_{FP}^{\theta_j^{cand}} = \frac{1}{kol_{cal,FP}} \sum_{k_{cal}: X_{k_{cal}}^{cal} \in \in FP} I \left(\tau_{(B_i),TP,k_{cal}}^{cal_{\mathfrak{D}_1}} < \theta_j^{cand} \right), \quad (40)$$

а в отрицательной ветви полнота выявления FN составляет:

$$Recall_{FN}^{\theta_j^{cand}} = \frac{1}{kol_{cal,FN}} \sum_{k_{cal}: X_{k_{cal}}^{cal} \in \in FN} I \left(\tau_{(B_i),FN,k_{cal}}^{cal_{\mathfrak{D}_0}} > \theta_j^{cand} \right), \quad (41)$$

Ограничения по правильной группе не характеризуют полный размер формируемой подозрительной области. Ее доля среди всех решений соответствующей ветви рассчитывается как:

$$SuspShare_{\mathfrak{D}_1}^{\theta_j^{cand}} = \frac{1}{kol_{cal,TP} + kol_{cal,FP}} \sum_{k_{cal}: X_{k_{cal}}^{cal} \in \in FP \cup TP} I \left(\tau_{(B_i),TP,k_{cal}}^{cal_{\mathfrak{D}_1}} < \theta_j^{cand} \right), \quad (42)$$

$$SuspShare_{\mathfrak{D}_0}^{\theta_j^{cand}} = \frac{1}{kol_{cal,TN} + kol_{cal,FN}} \sum_{k_{cal}: X_{k_{cal}}^{cal} \in \in FN \cup TN} I \left(\tau_{(B_i),FN,k_{cal}}^{cal_{\mathfrak{D}_0}} > \theta_j^{cand} \right), \quad (43)$$

Для каждой ветви выбирается допустимый порог с наибольшим $Recall$. При одинаковой полноте предпочтение получает кандидат с меньшей долей подозрительных объектов $SuspShare$. Если оба показателя совпадают, для $\mathfrak{D}_1$ выбирается наименьший порог, а для $\mathfrak{D}_0$ — наибольший. Такой порядок сохраняет достигнутое выявление ошибок при наименьшем объеме дополнительной проверки.

Полученные пороги θ_{TP} и θ_{FN} связывают оценки двух правил с диагностическими состояниями. Для

$$FPR_{TP}^{\theta_j^{cand}} = \frac{1}{kol_{cal,TP}} \sum_{k_{cal}: X_{k_{cal}}^{cal} \in TP} I \left(\tau_{(B_i),TP,k_{cal}}^{cal_{\mathfrak{D}_1}} < \theta_j^{cand} \right), \quad (37)$$

а для отрицательной ветви:

поступающего объекта применяемая ветвь определяется прогнозом ML-модели:

$$\begin{aligned} \tau_{(B_i)}(t) &= g_{\hat{y}_{(B_i)}(t)}(\mathcal{M}_{(B_i)}(t)) = \\ &= \begin{cases} \mathfrak{D}_1(\mathcal{M}_{(B_i)}(t)), & \hat{y}_{(B_i)}(t) = 1, \\ \mathfrak{D}_0(\mathcal{M}_{(B_i)}(t)), & \hat{y}_{(B_i)}(t) = 0. \end{cases} \end{aligned} \quad (44)$$

Выход $\mathfrak{D}_1$ образует оценку $\tau_{(B_i)}^{TP}$, а выход $\mathfrak{D}_0$ образует оценку $\tau_{(B_i)}^{FN}$. Их сопоставление с порогами определяет диагностическое состояние объекта:

$$\begin{aligned} Rep_{(B_i)}(t) &= \\ &= \begin{cases} \text{угроза,} & \hat{y}_{(B_i)}(t) = 1 \wedge \tau_{(B_i)}^{TP}(t) \geq \theta_{TP}, \\ \text{подозрительное,} & \hat{y}_{(B_i)}(t) = 1 \wedge \tau_{(B_i)}^{TP}(t) < \theta_{TP}, \\ \text{подозрительное,} & \hat{y}_{(B_i)}(t) = 0 \wedge \tau_{(B_i)}^{FN}(t) > \theta_{FN}, \\ \text{валидное,} & \hat{y}_{(B_i)}(t) = 0 \wedge \tau_{(B_i)}^{FN}(t) \leq \theta_{FN}. \end{cases} \end{aligned} \quad (45)$$

Таким образом, правило $\mathfrak{D}_1$ оценивает соответствие положительного прогноза профилям правильно выявленных угроз, а правило $\mathfrak{D}_0$ устанавливает сходство отрицательного прогноза с известными пропусками. Сопоставление полученной оценки с калибровочным порогом определяет диагностическое состояние объекта при заданном ограничении на долю правильных решений, направляемых на дополнительную проверку.

Экспериментальное исследование метода

Экспериментальная проверка была построена по двум следствиям выдвинутой гипотезы. Если диагностический профиль содержит сведения о правильности решения ML-модели, правила $\mathfrak{D}_1$ и $\mathfrak{D}_0$ должны сохранять различимость соответствующих групп на независимых данных. После калибровки порогов выявленные различия должны проявляться в составе диагностических состояний, обеспечивая дополнительный охват ошибочных решений при ограниченной доле правильных объектов, направляемых на проверку. В соответствии с этим сначала оценивалась различимость профилей, а затем охват ранее не наблюдавшейся вредоносной активности и связанная с ним нагрузка.

Исследование проведено на девяти самостоятельных задачах. Шесть из них были сформированы по сетевым данным CIC-IDS2017³ и UNSW-NB15⁴, три использовали табличные признаки смарт-контрактов, исход-

ный код Solidity⁵ и байткод⁶ [16, 17]. Feature Ablation, GradientSHAP и LIME были выбраны как методы, основанные соответственно на проверочном изменении признаков, градиентном приближении их вкладов и локальной суррогатной модели. Рассматривались три отдельных объяснителя, две парные конфигурации и совместное применение трех методов. Такой состав позволил проверить, зависит ли диагностическая информация от принципа формирования атрибуций и их совместного использования.

Данные, использованные для обучения ML-модели, формирования диагностических правил, калибровки порогов и итоговой оценки, не пересекались. Эталонные группы и правила $\mathfrak{D}_1$, $\mathfrak{D}_0$ формировались по D_{fit} , а пороги определялись по D_{cal} . Для сопоставимости XAI-конфигураций во всех задачах принималось $\alpha_{TP} = \alpha_{TN} = 0,1$. В семи из девяти задач один или несколько типов вредоносной активности исключались из всех предшествующих частей данных и вводились только в D_{test} . Наборы BCCC и Solidity не содержат самостоятельных типов уязвимостей, поэтому на них оценивался перенос диагностических правил без моделирования появления нового типа угрозы.

Способность правил различать правильные и ошибочные решения оценивалась по AUC. Для $\mathfrak{D}_1$ сопоставлялись оценки объектов TP и FP, для $\mathfrak{D}_0$ объектов TN и FN. AUC рассчитывалась на D_{test} отдельно для шести XAI-конфигураций. В табл. 1 приведены минимальные и максимальные значения AUC среди шести XAI-конфигураций, характеризующие зависимость разделяющей способности правил от способа получения локального объяснения.

Таблица 1
Различение правильных и ошибочных решений на тестовых данных

Сценарий	$AUC(\mathfrak{D}_1)$	$AUC(\mathfrak{D}_0)$
CIC-IDS2017: DDoS	0,891—0,968	0,494—0,516
CIC-IDS2017: PortScan	0,940—0,996	0,504—0,654
CIC-IDS2017: DoS GoldenEye, FTP-Patator, SSH-Patator	0,894—0,925	0,525—0,638
UNSW-NB15: Fuzzers	0,714—0,767	0,766—0,789
UNSW-NB15: Generic	0,861—0,885	0,909—0,916
UNSW-NB15: Exploits, DoS, Reconnaissance	0,741—0,763	0,890—0,904
Смарт-контракты: табличные признаки	0,606—0,639	0,575—0,598
Смарт-контракты: Solidity	0,731—0,766	0,579—0,673
Смарт-контракты: байткод	0,674—0,717	0,717—0,759

³ Network Intrusion dataset (CIC-IDS-2017) // Kaggle. URL: <https://www.kaggle.com/datasets/chethuhn/network-intrusion-dataset> (дата обращения: 14.08.2026); Sharafaldin I., Lashkari A.H., Ghorbani A.A. Toward generating a new intrusion detection dataset and intrusion traffic characterization // ICISp. 2018. Vol. 1. Pp. 108—116.

⁴ UNSW_NB15 // Kaggle. URL: <https://www.kaggle.com/datasets/mrwellsdavid/unswnb15> (дата обращения: 14.08.2026); Moustafa N., Slay J. UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set) // 2015 military communications and information systems conference (MilCIS). IEEE, 2015. Pp. 1—6.

⁵ Vulnerable Solidity Smart Contracts dataset // Kaggle. URL: <https://www.kaggle.com/datasets/rha4365/vulnerable-solidity-smart-contracts-dataset> (дата обращения: 14.08.2026).

⁶ Vulnerable Smart Contracts Detection using ML/DL (BCCC-VolSCs-2023) // Kaggle. URL: <https://www.kaggle.com/datasets/bcccdatasets/bccc-vulscs-2023> (дата обращения: 14.08.2026).

Результаты выявили различный характер диагностической информации в двух ветвях прогноза. Правило $\mathfrak{D}_1$ различало TP и FP во всех девяти задачах и при каждой XAI-конфигурации. Для $\mathfrak{D}_0$ наиболее выраженное разделение FN и TN получено в задачах UNSW-NB15. В сценариях CIC-IDS2017 результат зависел от сочетания данных и объяснителя. Для PortScan смена XAI-конфигурации заметно изменяла различимость, а в DDoS значения оставались около уровня случайного упорядочения. При анализе смарт-контрактов диагностическая информация сохранялась во всех трех представлениях данных, хотя значения AUC были ниже, чем в большинстве сетевых задач. Следовательно, в положительной ветви различия между профилями правильных и ошибочных решений сохранялись во всех задачах. В отрицательной ветви возможность их разделения определялась сочетанием данных, ML-модели и XAI-конфигурации, что подтверждает необходимость самостоятельного формирования $\mathfrak{D}_1$ и $\mathfrak{D}_0$.

AUC характеризует упорядочение правильных и ошибочных решений, но не показывает результат применения калиброванных порогов. Чтобы установить, приводит ли выявленная различимость к дополнительному охвату ранее не наблюдавшейся вредоносной активности, рассматривалась целевая группа $\Omega_{\text{цель}}$ представленная только в D_{test} . Исходный охват этой группы определялся долей положительных прогнозов ML-модели:

$$\rho_{ML} = \frac{1}{|\Omega_{\text{цель}}|} \sum_{j \in \Omega_{\text{цель}}} I(\hat{y}_j = 1), \quad (46)$$

После применения диагностических правил в области наблюдения сохранялись объекты, получившие состояния «угроза» и «подозрительное». Их совокупная доля составляла:

$$\rho_{ML+XAI} = \frac{1}{|\Omega_{\text{цель}}|} \sum_{j \in \Omega_{\text{цель}}} I(ALERT_j). \quad (47)$$

Разность $\rho_{ML+XAI} - \rho_{ML}$ определяет дополнительную долю удержанной вредоносной активности, переведенной правилом $\mathfrak{D}_0$ в подозрительное состояние. Однако этот результат не характеризует объем дополнительной проверки, поскольку целевая группа составляет лишь часть подозрительных объектов. Доля

всех объектов D_{test} , получивших подозрительное состояние, определялась выражением:

$$\rho_{\text{доп}} = \frac{1}{N_{\text{test}}} \sum_{j=1}^{N_{\text{test}}} I(Rep_j = \text{подозрительное}), \quad (48)$$

где N_{test} — число объектов тестовой выборки. Воздействие метода на обработку объектов нормы оценивалось показателем:

$$\rho_{\text{норм}} = \frac{\sum_{j=1}^{N_{\text{test}}} I(y_j=0 \wedge Rep_j=\text{подозрительное})}{\sum_{j=1}^{N_{\text{test}}} I(y_j=0)}. \quad (49)$$

Объем дополнительной проверки не показывает, насколько избирательно правило $\mathfrak{D}_0$ выделяет пропуски угроз. Поэтому доля FN среди подозрительных объектов отрицательной ветви сопоставлялась с их долей среди всех отрицательных прогнозов ML-модели. Отношение этих величин определялось коэффициентом:

$$\begin{aligned} Lift_{FN} &= \frac{\sum_{j=1}^{N_{\text{test}}} I(\hat{y}_j=0 \wedge y_j=1 \wedge Rep_j=\text{подозрительное})}{\sum_{j=1}^{N_{\text{test}}} I(\hat{y}_j=0 \wedge Rep_j=\text{подозрительное})} \\ &= \frac{\sum_{j=1}^{N_{\text{test}}} I(\hat{y}_j=0 \wedge y_j=1)}{\sum_{j=1}^{N_{\text{test}}} I(\hat{y}_j=0)}. \end{aligned} \quad (50)$$

$Lift_{FN} > 1$ означает повышенную концентрацию FN в подозрительной области. В табл. 2 исходный охват приведен одним значением, а результаты шести XAI-конфигураций — диапазонами.

Во всех семи сценариях итоговый охват удержанной вредоносной активности превышал исходный результат ML-модели. Возможный прирост ограничивался долей целевых объектов, пропущенных исходной ML-моделью. Его фактическое значение зависело от разделяющей способности $\mathfrak{D}_0$ и выбранного порога. Поэтому почти полный исходный охват Generic оставлял небольшой резерв для улучшения. В DDoS значительная часть целевой активности оставалась пропущенной, однако значения AUC, близкие к 0,5, сопровождалась малым приростом. Для Fuzzers правило $\mathfrak{D}_0$ сохраняло высокую разделяющую способность при всех XAI-конфигурациях, тогда как результат PortScan заметно зависел от выбранного объяснителя.

Таблица 2

Охват удержанной вредоносной активности и показатели дополнительной проверки

Сценарий	ρ_{ML}	ρ_{ML+XAI}	$Lift_{FN}$	$\rho_{\text{норм}}$	$\rho_{\text{доп}}$
DDoS	0,6332	0,636—0,647	1,134—1,789	0,0253—0,0413	0,0202—0,1280
DoS GoldenEye, FTP-Patator, SSH-Patator	0,2629	0,3210—0,4757	1,384—1,643	0,0345—0,0625	0,0914—0,2303
PortScan	0,0030	0,0389—0,3464	1,120—1,496	0,0312—0,0452	0,0322—0,1775
Exploits, DoS, Reconnaissance	0,8105	0,8406—0,9242	2,318—2,974	0,0290—0,1193	0,0743—0,1834
Fuzzers	0,2370	0,4404—0,5822	1,494—1,536	0,0673—0,1165	0,2052—0,2876
Generic	0,9782	0,9839—0,9920	3,529—3,949	0,1032—0,1273	0,0731—0,0919
Смарт-контракты: байткод	0,3750	0,4571—0,5571	1,928—2,190	0,0410—0,1270	0,0571—0,1700

Значения $Lift_{FN} > 1$, полученные во всех сценариях, свидетельствуют о преимущественной концентрации пропусков угроз в подозрительной области отрицательной ветви. Этот показатель дополняет оценку охвата, но не заменяет ее. Для Generis высокая концентрация FN сочеталась с небольшим приростом из-за малого числа исходных пропусков, тогда как для PortScan меньшие значения $Lift_{FN}$ могли сопровождаться существенно большим дополнительным охватом.

Увеличение охвата обычно расширяло область дополнительной проверки $P_{доп}$, однако эта зависимость не была пропорциональной. В DDoS изменение $P_{доп}$ почти не влияло на результат, а в PortScan и объединенном сценарии CIC-IDS2017 отдельные XAI-конфигурации обеспечивали заметный прирост при ограниченном воздействии на объекты нормы. Для Fuzzers больший охват сопровождался более широкой проверкой, тогда как в Generis воздействие на норму оказалось высоким относительно полученного прироста.

Полученные зависимости показывают, что XAI-конфигурация определяет соотношение между охватом пропущенной активности, концентрацией FN и объемом дополнительной проверки. Универсально предпочтительного варианта по совокупности этих характеристик не выявлено. Более широкий охват мог сопровождаться увеличением очереди, а высокая концентрация FN могла достигаться при охвате лишь небольшой части исходных пропусков. Эксперимент подтвердил возможность сохранять часть ранее не наблюдавшейся вредоносной активности в области наблюдения, но показал, что практический результат метода нельзя характеризовать максимумом одного показателя. Эта зависимость определяет требование к последующей методике применения, предусматрива-

ющей выбор XAI-конфигурации по допустимому соотношению охвата, концентрации FN и объема дополнительной проверки.

Заключение

Исследование показало, что количественное описание локального объяснения содержит диагностическую информацию о правильности решения ML-модели. В положительной и отрицательной ветвях прогноза эта информация проявляется по-разному, что обосновывает формирование самостоятельных правил $\mathfrak{D}_1$ и $\mathfrak{D}_0$. Их применение с калиброванными порогами дополняет исходный прогноз подозрительным состоянием без изменения классификатора.

Экспериментальное исследование подтвердило гипотезу в положительной ветви всех девяти задач при каждой XAI-конфигурации. В отрицательной ветви возможность разделения TN и FN зависела от данных и способа формирования объяснения. Во всех семи сценариях с удержанной вредоносной активностью часть пропущенных объектов была переведена в подозрительное состояние, где концентрация FN превышала исходную концентрацию среди отрицательных прогнозов. Следовательно, дополнительный охват обеспечивался не только расширением области проверки, но и преимущественным выделением пропусков. Соотношение достигнутого результата и объема проверки различалось между XAI-конфигурациями, поэтому универсально предпочтительный вариант не установлен. Полученная зависимость определяет необходимость разработки методики выбора ML- и XAI-конфигураций, допустимых значений α и условий актуализации диагностических правил.

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. Скиба В.Ю., Балябин А.А., Петренко С.А. Проблема обеспечения киберустойчивости облачных платформ управления киберфизическими объектами // Правовая информатика. 2025. № 4. С. 26—38.
3. 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.
4. Hasan M. et al. Detecting anomalies in blockchain transactions using machine learning classifiers and explainability analysis // Blockchain: Research and Applications. Elsevier, 2024. Vol. 5. No. 3. Article 100207.
5. Hosain Y., Çakmak M. XAI-XGBoost: an innovative explainable intrusion detection approach for securing internet of medical things systems // Scientific reports. Nature Publishing Group London, 2025. Vol. 15. No. 1. Article 22278.
6. Krishnan D., Singh S., Sugumaran V. Explainable AI for zero-day attack detection in IoT networks using attention fusion model // Discov. Internet Things. 2025. Vol. 5. No. 1. Article 83.
7. Kaya K. et al. X-cba: Explainability aided catboosted anomaly intrusion detection system // ICC 2024-IEEE International Conference on Communications. IEEE, 2024. Pp. 2288—2293.

8. 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. Pp. 1—6.
9. 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. Frontiers Media SA, 2025. Vol. 6. Article 1697204.
10. Mougan C. et al. Explanation Shift: How Did the Distribution Shift Impact the Model? // Transactions on Machine Learning Research. 2024.
11. 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. Pp. 550—561.
12. Nauta M. et al. From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI // ACM Comput. Surv. 2023. Vol. 55. No. 13s. Pp. 1—42.
13. Hedström A. et al. Quantus: An Explainable AI Toolkit for Responsible Evaluation of Neural Network Explanations and Beyond // Journal of Machine Learning Research. 2023. Vol. 24. No. 34. Pp. 1—11.
14. Weber R.O. et al. XAI is in trouble // AI Magazine. 2024. Vol. 45. No. 3. Pp. 300—316.
15. Лисов Д.Н. Интерпретационная модель наблюдения и диагностирования состояния элементов блокчейн-систем по решениям моделей машинного обучения // Правовая информатика. 2026. № 3. В печати.
16. Alnuman R. et al. Ensemble multi-label machine learning solidity smart contract vulnerability detection model // Cluster Comput. 2025. Vol. 28. No. 16. Article 1014.
17. HajiHosseiniKhani S., Lashkari A.H., Oskui A.M. Unveiling vulnerable smart contracts: Toward profiling vulnerable smart contracts using genetic algorithm and generating benchmark dataset // Blockchain: Research and Applications. Elsevier, 2024. Vol. 5. No. 1. Article 100171.

SECTION:

SYSTEM ANALYSIS, MANAGEMENT, ARTIFICIAL INTELLIGENCE, AND INFORMATION PROTECTION

A METHOD FOR INTERPRETATIVE DIAGNOSIS OF THE STATE OF BLOCKCHAIN SYSTEM COMPONENTS USING MACHINE LEARNING MODEL SOLUTIONS

Lisov D.N.⁷, Gnidko K.O.⁸

Keywords: *explainable artificial intelligence, explanation quality assessment, local explanation, statistical learning, diagnostic rules, diagnostic profile, threshold calibration, suspicious state, new types of threats.*

Abstract

Purpose of work: *to develop a method for interpretative diagnosis of the state of blockchain system components based on machine learning model outputs when new types of threats emerge.*

Research method: *the differences between correct and incorrect decisions made by the ML model were investigated using local explanation metrics. These metrics characterised changes in the model's output following controlled substitution of input features, the structure of attributions, and the consistency of results across multiple explainers. The resulting diagnostic profiles were used to formulate rules using statistical learning methods. Threshold limits were determined in accordance with statistical decision theory, whilst limiting the proportion of correct predictions sent for further verification. The distinguishability between correct and incorrect decisions was assessed using the AUC on independent data. The practical outcome was characterised by a change in the coverage of previously unobserved malicious activity, the concentration of missed threats in the suspicious region, and the volume of additional verification.*

⁷ **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

⁸ **Konstantin O. Gnidko**, Dr.Sc. (Technology), Associate Professor, Deputy Head of Group at the Scientific Center for Information Technology and Artificial Intelligence, Sirius University of Science and Technology, Sirius, Krasnodar region, Russian Federation. ORCID: 0000-0002-8605-8865.

E-mail: gnidko.ko@talantiuspeh.ru

Results of the study: a method has been developed that supplements the binary prediction of an ML model with a diagnostic state generated on the basis of local explanation characteristics, without altering the original classifier. The scientific novelty lies in the construction of separate diagnostic rules for the positive and negative branches of the prediction and in the independent calibration of their thresholds, taking into account the acceptable proportion of correct decisions sent for additional verification. An experimental study using network data and various representations of smart contracts has shown that the generated rules distinguish between correct and incorrect decisions across different XAI configurations. In scenarios involving malicious activity not encountered during the training of the ML model or the development of diagnostic rules, some of the missed objects were reclassified as suspicious. At the same time, the resulting volume of additional verification was determined. The results confirmed the diagnostic informativeness of local explanation metrics and demonstrated its dependence on the data, the ML model and the XAI configuration. No universally preferred XAI configuration has been identified.

References

1. Larina M.V., Skiba V.Iu. Model' tsifrovoy platformy dlia analiza ustoichivosti k kvantovym ugrozam // Pravovaia informatika. 2025. No. 3. S. 12—26.
2. Skiba V.Iu., Baliabin A.A., Petrenko S.A. Problema obespecheniia kiberustoiichivosti oblachnykh platform upravleniia kiberfizicheskimi ob'ektami // Pravovaia informatika. 2025. No. 4. S. 26—38.
3. Petrenko S.A., Sundeev P.V. Kontrol' trebovaniia k zashchite informatsii na osnove modeli logicheskogo vyvoda i grafovoi neironnoi seti // Pravovaia informatika. 2026. No. 2. S. 85—94.
4. Hasan M. et al. Detecting anomalies in blockchain transactions using machine learning classifiers and explainability analysis // Blockchain: Research and Applications. Elsevier, 2024. Vol. 5. No. 3. Article 100207.
5. 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. Vol. 15. No. 1. Article 22278.
6. Krishnan D., Singh S., Sugumaran V. Explainable AI for zero-day attack detection in IoT networks using attention fusion model // Discov. Internet Things. 2025. Vol. 5. No. 1. Article 83.
7. Kaya K. et al. X-cba: Explainability aided catboosted anomal-e for intrusion detection system // ICC 2024-IEEE International Conference on Communications. IEEE, 2024. Pp. 2288—2293.
8. 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. Pp. 1—6.
9. 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. Frontiers Media SA, 2025. Vol. 6. Article 1697204.
10. Mougan C. et al. Explanation Shift: How Did the Distribution Shift Impact the Model? // Transactions on Machine Learning Research. 2024.
11. 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. Pp. 550—561.
12. Nauta M. et al. From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI // ACM Comput. Surv. 2023. Vol. 55. No. 13s. Pp. 1—42.
13. Hedström A. et al. Quantus: An Explainable AI Toolkit for Responsible Evaluation of Neural Network Explanations and Beyond // Journal of Machine Learning Research. 2023. Vol. 24. No. 34. Pp. 1—11.
14. Weber R.O. et al. XAI is in trouble // AI Magazine. 2024. Vol. 45. No. 3. Pp. 300—316.
15. Lisov D.N. Interpretatsionnaia model' nabliudeniia i diagnostirovaniia sostoianiia elementov blokchein-sistem po resheniiam modelei mashinnogo obucheniia // Pravovaia informatika. 2026. No. 3. V pechaty.
16. Alnuman R. et al. Ensemble multi-label machine learning solidity smart contract vulnerability detection model // Cluster Comput. 2025. Vol. 28. No. 16. Article 1014.
17. HajiHosseinKhani S., Lashkari A.H., Oskui A.M. Unveiling vulnerable smart contracts: Toward profiling vulnerable smart contracts using genetic algorithm and generating benchmark dataset // Blockchain: Research and Applications. Elsevier, 2024. Vol. 5. No. 1. Article 100171.

