ИНЖИНИРИНГ ОНТОЛОГИЙ

УДК 004.056 *Научная статья* DOI: 10.18287/2223-9537-2025-15-1-96-113

Обнаружение и объяснение аномалий в индустриальных системах Интернета вещей на основе автокодировщика

© 2025, Д.А. Левшун, Д.С. Левшун, И.В. Котенко ✉

Санкт-Петербургский Федеральный исследовательский центр Российской академии наук (СПб ФИЦ РАН), Санкт-Петербург, Россия

Аннотация

В индустриальных системах Интернета вещей объяснение аномалий может помочь выявить узкие места и способствовать оптимизации процессов. В статье предлагается подход к обнаружению аномалий при помощи автокодировщика и их объяснению на основе метода аддитивных объяснений Шепли. Результатом объяснения аномалий является предоставление набора признаков данных в индустриальных системах Интернета вещей, более всего повлиявших на обнаружение аномальных случаев. Новизна предложенного подхода заключается в способности определять вклад отдельных признаков для выбранных образцов данных и вычислять средний вклад для всей выборки в виде рейтинга признаков. Оценка предлагаемого подхода проводится на наборах данных индустриального Интернета вещей с различным количеством признаков и объёмом данных. Итоговая *F*-мера обнаружения аномалий достигает 88-93%, что превосходит рассмотренные в статье аналоги. Показано, как объяснимый искусственный интеллект может помочь раскрыть причины аномалий в отдельных образцах и в выборке данных. В качестве теоретической значимости предложенного подхода можно выделить то, что анализ аномалий помогает разобраться в работе интеллектуальных моделей обнаружения, позволяя выявлять факторы, влияющие на их выводы, и открывая ранее незамеченные закономерности. На практике предложенный метод может улучшить понимание текущих процессов для операторов систем безопасности, способствуя выявлению угроз и обнаружению ошибок в данных.

Ключевые слова: *информационная безопасность, обнаружение аномалий, индустриальные системы Интернета вещей, автокодировщик, объяснимый искусственный интеллект*.

Цитирование: *Левшун Д.А., Левшун Д.С., Котенко И.В.* Обнаружение и объяснение аномалий в индустриальных системах Интернета вещей на основе автокодировщика. *Онтология проектирования*. 2025. Т.15, №1(55). С.96-113. DOI:10.18287/2223-9537-2025-15-1-96-113.

Финансирование: исследование выполнено за счёт гранта Санкт-Петербургского научного фонда № 23-РБ-01-09.

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Введение

Технологии Интернета вещей (умного города, умных зданий, заводов, ферм и др.) призваны осуществлять сбор и анализ данных со всех объектов инфраструктуры, контролировать их работу и управлять ими [1, 2]. В промышленной области индустриальные системы Интернета вещей (ИСИВ) позволяют увеличить эффективность производственных процессов, снизить затраты и обеспечить более высокую степень автоматизации. Подобные системы представляют собой сеть взаимосвязанных устройств (датчиков и систем), используемых для сбора, передачи и анализа данных.

В то время как экономическая выгода от интеллектуализации очевидна, обратная сторона этого процесса заключается в значительном увеличении ущерба, который может быть при-

чинён посредством информационных атак[1]. Процесс интеллектуализации объектов критически важной инфраструктуры далёк от завершения, а решения по защите подобных объектов не успевают за темпом их технологического развития. Это создаёт разрыв между возможностями атакующих и защитных мер.

В то же время существует большое количество средств защиты информации, основанных на методах искусственного интеллекта (ИИ) [3, 4]. Большинство передовых технологий, таких как глубокие нейронные сети (НС), работают в режиме «чёрного ящика», когда причины принятия решения относительно состояния безопасности остаются неизвестными. В данном случае большим подспорьем для операторов систем безопасности становится объяснимый ИИ (ОИИ) [5]. Объяснение отклонений в данных ИСИВ позволяет различить выбросы[2] и аномалии[3], специфичные для этой предметной области. Механизм обнаружения аномалий с использованием ОИИ способен выделить как аномалии, заслуживающие внимания с точки зрения поведения системы, так и выбросы из-за редкости значений в данных, не интересные с точки зрения её безопасности. Определение причин отклонений может сократить объём проверки, выполняемой экспертами. ОИИ может использовать различные методы, чтобы предоставить информацию о том, какие факторы способствовали возникновению аномалии.

В этой статье предложен подход к обнаружению аномалий при помощи автокодировщика (АК)[4] и их объяснению на основе метода аддитивных объяснений Шепли (*SHapley Additive exPlanations*, *SHAP*) [6]. Модель АК используется для реконструкции входных данных от датчиков ИСИВ. Целью объяснения аномалий является предоставление набора признаков, которые способствуют ошибке реконструкции аномальных случаев.

1 Методы обнаружения аномалий

1.1 Обнаружение аномалий на основе интеллектуальных методов

Как правило, обнаружение атак осуществляется системами обнаружения вторжений, которые используют известные сигнатуры атак и ищут аномалии в виде отклонений от нормального поведения. Масштаб и разнообразие данных часто приводят к тому, что создавать «ручные» правила обнаружения атак и уязвимостей становится непрактичным. Использование машинного обучения позволяет искать закономерности в больших наборах данных и обучаться на них, чтобы предотвратить аналогичные атаки и динамически реагировать на изменение поведения ИСИВ.

Среди методов машинного обучения для обнаружения атак в ИСИВ часто применяются методы классификации: метод *k*-ближайших соседей [7], метод опорных векторов [8], скрытые марковские модели [9] и др. Традиционное машинное обучение зависит от экспертов, которые создают иерархию признаков данных. Для глубокого машинного обучения не всегда требуется наличие размеченного набора данных, оно способно использовать неструктурированные данные и автоматически определять особенности, которые отличают один образец от других. По этой причине широкое распространение получили глубокие НС для обнаружения атак [10, 11]. Сравнительно простым вариантом такой архитектуры являются глубокие НС с

[1] *Информационная атака* — это преднамеренное действие, направленное на нарушение работы компьютерных систем, сетей или цифровых устройств с целью нанесения ущерба.

[2] *Выброс* – в статистике результат измерения, выделяющийся из общей выборки.

[3] *Аномалия* – это отклонение поведения системы от стандартного (ожидаемого). Они могут включать выбросы, а также охватывать более широкий спектр несоответствий, таких как временные изменения или неожиданные тенденции. Все выбросы могут быть аномалиями, но не все аномалии являются выбросами.

[4] *Автокодировщик* (англ. *Autoencoder, AE*) – специальная архитектура искусственных НС, состоящая из двух частей: кодировщика (англ. *encoder*) и декодировщика (англ. *decoder*).

прямой связью для обнаружения вторжений. Но таким моделям часто не хватает способности обучаться на предыдущих входных данных и предыдущих итерациях обучения. Свёрточные НС (*Convolutional Neural Network, CNN*) позволяют анализировать структурированные данные [12]. Временные зависимости помогают учитывать рекуррентные НС, в частности с блоками долговременной краткосрочной памяти (*Long Short-Term Memory, LSTM*) [13].

При обучении моделей обнаружения аномалий получение большого количества маркированных аномальных данных, как это требуется при обучении с учителем, может быть трудоёмким, поскольку требуется ручная работа эксперта в предметной области. В результате полностью контролируемое обнаружение аномалий часто непрактично [14]. Ряд исследований направлен на применение методов обучения без учителя, для которых не требуются предварительные знаний о вредоносной активности.

Известной моделью подобного типа является АК, который представляет собой глубокую НС для реконструкции входа (входных данных). АК формирует данные в более низкой размерности (кодирование) и реконструирует данные в исходной размерности (декодирование), т.е. восстанавливает оригинальные данные. Ошибка реконструкции – это мера, которая используется для оценки качества работы АК (и других моделей, которые могут восстанавливать данные). Ошибка реконструкции определяется как разница между оригинальными и восстановленными данными.

Для обнаружения аномалий АК анализирует функцию идентичности нормальных экземпляров, которые соответствуют ожидаемым или типичным образцам. Эти данные представляют собой поведение или характеристики, которые модель считает стандартными, и на основе которых она будет определять, что является аномалией. Аномалии имеют высокую погрешность (ошибку) реконструкции, что и способствует их обнаружению. На таких принципах работают инструменты *RANSynCoder* [15] и *InterFusion* [16]. В [17] проводится сравнение ряда АК, включая вариационный АК (*Variational AE, VAE*) и неполный АК с *CNN*-1D с методом главных компонент обнаружения аномальной активности в ИСИВ.

В качестве кодера и декодера также могут использоваться рекуррентные НС, что даёт преимущество при анализе временных рядов. Например, такой подход реализован в *LSTM-FWED* (*LSTM Encoder-Decoder Feature Weight*) [18]. АК может быть включён в генеративно-состязательную сеть (*Generative Adversarial Network, GAN*), как это сделано в *USAD* [19], или объединён с графовой НС, как в *FuSAGNet* [20].

1.2 Объяснимое обнаружение аномалий

Система ОИИ пытается описать своё поведение, чтобы сделать его более понятным для людей. ОИИ позволяет исследователям и разработчикам анализировать, какие факторы влияют на результаты работы моделей.

Интерпретируемость модели можно разделить на две категории: глобальную и локальную. Глобальная интерпретируемость означает, что пользователи могут понять модель непосредственно из её общей структуры. Локальная интерпретируемость проверяет входные данные и пытается выяснить, почему модель принимает определённое решение. Примерами глобальных моделей ОИИ можно назвать нейронно-аддитивную модель (*Neural Additive Model, NAM*) [21] и объяснимую сеть глубокого доверия (*Deep Belief Networks, DBN*) [22]. Примерами локальных моделей являются *TRUST* (*Transparency Relying Upon Statistical Theory*) [23], *LIME* (*Local Interpretable Model-Agnostic*) [24] и *SHAP* [6]. Ряд моделей совместно используют глобальную и локальную интерпретации, например, [25, 26].

Методы ОИИ подразделяется на методы апостериорного объяснения и методы предварительного объяснения. Методы апостериорного объяснения используются для объяснения производительности модели после её обучения [27, 28]. Методы предварительного объясне-

ния используются для объяснения производительности модели до её обучения, например, *LIME*. Данная модель аппроксимирует предсказания моделей чёрного ящика и обучается на локальной суррогатной модели для интерпретации отдельных прогнозов.

Большинство исследований об объяснениях вредоносной активности посвящены моделям, обучаемым с учителем, т.е. на заранее размеченных данных. Так, для обнаружения атак в [29] модели глубокого обучения объединяются с *SHAP*, который позволяет оценить вклад каждого признака в предсказание модели на основе теории игр. В [30] предложен подход к объяснению конкретной аномалии путём создания для неё случайного леса и дальнейшего извлечения правил, которые объясняют классификацию этой аномалии.

Модели для обнаружения аномалий, как правило, основаны на обучении без учителя. Объяснение результатов подобных моделей встречается реже, чем для моделей, основанных на обучении с учителем [31]. В [32] проводится анализ градиентов, «вносимых» каждым признаком конкретного экземпляра данных, которые можно получить из вариационного АК посредством автоматического дифференцирования[5]. Объяснимое обнаружение аномалий в [33] осуществляется на основе метода главных компонент и векторов Шепли, которые используются для объяснения полученных ошибок реконструкции. Схожим образом в [31, 34-36] объясняются при помощи метода *SHAP* ошибки реконструкции, полученные АК для выявления аномалий. Перечисленные подходы позволяют выделить признаки данных, которые наиболее важны с точки зрения их влияния на ошибку реконструкции для конкретного экземпляра данных.

2 Предлагаемый подход

Предлагается использовать АК в качестве модели обучения без учителя для обнаружения аномалий на основе полученных ошибок реконструкции данных. В обучении без учителя алгоритмы предназначены для выявления скрытых структур в данных без явных указаний на то, что искать. Ошибка реконструкции при этом представляет собой разницу между входным и выходным (реконструированным) значением. Экземпляры данных с ошибкой реконструкции, превышающий установленный для нормы порог, считаются аномальными. Целью модели ОИИ является определение признаков, оказывающих наибольшее влияние на результаты работы модели. Для этого модель ОИИ на основе *SHAP* вычисляет вектор Шепли для реконструированных признаков и связывает их с истинными значениями входных данных.

Предлагаемый подход включает следующие этапы.

2.1 Предобработка данных

После сбора данных для обучения модели необходимо провести их предобработку, включая удаление пропусков и дубликатов, нормализацию, кодирование категориальных признаков и разделение на обучающую и тестовую выборки. Подготовленные входные данные для модели ОИИ представляют собой набор векторов признаков (экземпляров):

$$X = \{x_1, x_2, \ldots, x_N\},$$
$$x_i = \{f_{i1}, f_{i2}, \ldots, f_{iM}\}$$

где x_i – i-ый вектор признаков, N – число векторов (длина набора данных), f_{ij} – значение j-того признака для i-ого вектора, M – число признаков данных.

Можно выделить множества значений отдельных признаков:

$$F_j = \{f_{1j}, f_{2j}, \ldots, f_{Nj}\}.$$

[5] *Автоматическое дифференцирование* – это способ вычисления производной функции, заданный программно. Этот вид дифференцирования опирается на правило дифференцирования сложной функции, представление функции в виде последовательности элементарных операций и перегрузке программных инструкций (функций, операторов).

2.2 Подготовка модели реконструкции данных

На данном этапе определяется модель АК, включая выбор архитектуры, определение слоёв, настройку функции потерь и оптимизатора. В данном подходе предлагается использовать АК с полносвязными слоями. Размер входного слоя кодера и выходного слоя декодера соответствуют числу признаков (M).

Выбираются и оптимизируются следующие гиперпараметры модели: количество скрытых слоёв кодера, количество нейронов на скрытых слоях, размер скрытого пространства, функция активации на скрытых слоях и функция активации на выходном слое. Для этого предлагается использовать алгоритм байесовской оптимизации [37]. Множество искомых гиперпараметров модели обозначается как множество λ, а пространство поиска этих гиперпараметров – как множество Λ, так что $\lambda \in \Lambda$. Функцию оптимизации β можно представить как β: $(X, \Lambda) \rightarrow \Phi_\lambda$, где Φ_λ – итоговая модель реконструкции данных с гиперпараметрами λ.

2.3 Обучение модели реконструкции данных

Обучение осуществляется на основе АК на нормальных данных. Кодировщик принимает входные данные и преобразует их в сжатое представление или латентное пространство. Декодировщик принимает сжатое представление и пытается восстановить исходные данные. Во время обучения АК минимизирует разницу между исходными и восстановленными данными, используя функцию потерь. Процесс реконструкции данных при помощи модели АК можно представить в следующем виде:

$$\Phi\left(x_i\right) = x_i' = \left\{f_{i1}', f_{i2}', \ldots, f_{iM}'\right\}, \; \Phi : X \rightarrow X',$$

где x_i – входной вектор данных, $x_i \in X$; x'_i – выходной (реконструированный) вектор данных, $x'_i \in X'$, X' – множество выходных векторов.

Можно выделить множества значений отдельных реконструированных признаков:

$$F_j' = \left\{f_{1j}', f_{2j}', \ldots, f_{Nj}'\right\}.$$

2.4 Вычисление пороговых значений для обнаружения аномалий

На этом этапе определяется ошибка реконструкции нормальных данных. В качестве показателя предлагается использовать значение среднеквадратичной ошибки реконструкции. Для каждого вектора данных:

$$\varepsilon_i = \left|x_i - x_i'\right| = \frac{1}{M}\sum_{j=1}^{M}\left(f_{ij} - f_{ij}'\right)^2, \quad i = 1..N.$$

Для отдельных признаков:

$$\xi_j = \left|F_j - F_j'\right| = \frac{1}{N}\sum_{i=1}^{N}\left(f_{ij} - f_{ij}'\right)^2, \quad j = 1..M.$$

Таким образом, для исходного набора данных получается множество ошибок реконструкции отдельных экземпляров $E = \{\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_N\}$ и ошибки реконструкции признаков данных $R = \{\xi_1, \xi_2, \ldots, \xi_M\}$, где $\varepsilon_i = \varepsilon(x_i, x'_i)$ – ошибка реконструкции для i-того вектора данных.

Для определения порогового значения на обучающих данных используется метод 90-го процентиля, чтобы смягчить влияние выбросов в нормальных данных: $P(E \leq \delta) \geq 0.9$, где $P(E)$ – вероятностная мера, задающая распределение E, δ – пороговое значение ошибки реконструкции.

2.5 Обнаружение аномалий на тестовых данных

На данном этапе проводится сопоставление полученных ошибок прогнозирования с пороговым значением для выявления аномальных выбросов:

$$y_i = 1, \text{ если } \varepsilon_i \geq \delta, \text{ иначе } 0, \qquad \varepsilon_i \in E,$$

где $y_i \in (0,1)$ – метка состояния безопасности для вектора признаков, значение 1 соответствует аномалии, а 0 – норме.

Для входного набора данных X результатом обнаружения является множество:

$$Y = \{y_1, y_2, \ldots, y_N\}.$$

Отдельно можно выделить подмножество аномальных экземпляров:

$$X^* = \{x_i \mid y_i = 1\}, \; i = 1..n,$$

где n – количество аномальных образцов.

2.6 Выбор признаков с наибольшими ошибками реконструкции

Для этого необходимо определить упорядоченное множество:

$$R' = \{\xi_{(1)}, \xi_{(2)}, \ldots, \xi_{(M)}\},$$

для которого:

$$\xi_{(1)} \geq \xi_{(2)} \geq \ldots \geq \xi_{(M)}, \quad \xi_{(j)} \in R.$$

Здесь индекс в скобках обозначает порядковый номер, где $\xi_{(1)}$ – значение наибольшей ошибки реконструкции.

Выбирается подмножество r, которое содержит наибольшую ошибку реконструкции для m признаков: $r = \{\xi_{(1)}, \xi_{(2)}, \ldots, \xi_{(m)}\}$, $m \leq M$, $r \subseteq R'$.

2.7 Вычисление векторов Шепли для выбранных признаков

Для объяснения модели реконструкции необходимо определить модель объяснения Ω. Чтобы объяснить единичный случай x_i, модель использует упрощённые входные данные z и отображение h, такое, что $x = h(z)$, $z \in \{0, 1\}$. Таким образом, модель использует собственное упрощённое представление входных данных.

Для каждого признака из множества r необходимо объяснить, какие признаки (кроме выбранного) привели к ошибке реконструкции. Влияние каждого признака на реконструкцию определяется значением Шепли, которое описывает, как распределить общий выигрыш (или полезность) между участниками кооперативной игры. Значение Шепли для игрока определяется как среднее значение его предельного вклада по всем возможным порядкам входа игроков в игру. В данном случае число упорядоченных игроков равно M (число признаков). Пусть s подмножество j первых игроков (признаков) в этом упорядочении. В качестве характеристической функции кооперативной игры выступает модель АК, обозначенная как Ф. Тогда вклад j-го игрока определяется как разница между Ф(s) и Ф($s\backslash j$), где Ф(s) – результат реконструкции по выбранному подмножеству признаков, а Ф($s\backslash j$) – результат реконструкции по выбранному подмножеству признаков, исключая j-ый признак. Предыдущие разности вычисляются для всех возможных подмножеств $s \subseteq F\backslash\{j\}$, где $F\backslash\{j\}$ – множество всех признаков, за исключением j-го.

Значение Шепли для j-го признака вектора данных x_i при этом вычисляется как [6]:

$$\varphi_{ij}(\Phi, x_i) = \sum_{s \subseteq F_i \backslash \{j\}} \frac{|s|!(M - |s| - 1)!}{M!} \left[\Phi(s) - \Phi(s \setminus j)\right],$$

где символом «!» обозначена функция факториала, а символом | | – размер множества.

Разница значений Ф(s) и Ф($s\backslash j$) определяет, как интересующий j-ый признак способствует реконструкции данных в заданном подмножестве признаков. Функцию модели ОИИ можно определить в форме линейной функции:

$$\Omega(z_i) = \varphi_{i0} + \sum_{j=1}^{M} \varphi_{ij} z_i,$$

где φ_{i0} – значение Шепли, при котором подмножество признаков является пустым.

Вектор Шепли – это вектор значений Шепли для всех игроков в игре. Для каждого признака в множестве r вычисляются векторы, которые содержат значения Шепли для всех признаков, помимо выбранного. Для этого исходная модель АК изменяется путём обнуления веса выбранного признака: Φ^k, $k = 1..m$. Так модель объяснения принимает на вход модель Φ^k и векторы данных x_i, предсказывает целевое значение k-го признака и определяет вектор Шепли. Результатом этого шага является матрица $\mathbf{V}$ (размером $m \times M$), в которой каждая строка представляет вектор Шепли для одного признака с наибольшей ошибкой реконструкции:

$$\mathbf{V}_i = \left(\varphi_{ij}(\Phi^k, x_i)\right)_{k=1,j=1}^{m,M},$$

Для каждого признака определяется среднее абсолютное значение Шепли. В результате получается вектор:

$$V_i = \{v_{i1}, v_{i2}, \ldots v_{iM}\},$$

$$v_{ij} = \frac{1}{m}\sum_{k=1}^{m}\left|\varphi_{ij}\left(\Phi^k, x_i\right)\right|, \quad j = 1..M.$$

Для выборки данных таким образом можно получить множество векторов:

$$\{V_1, V_2, \ldots, V_N\}.$$

2.8 Определение общего вклада признаков в обнаружение аномалий

Наивысшее значение Шепли соответствует наибольшему его вкладу в полученную ошибку реконструкции, т.е. впоследствии в результат обнаружения аномалий. Для каждого вектора V_i значения Шепли ранжируются от наибольшего к наименьшему:

$$v_{i(1)} \geq v_{i(2)} \geq \ldots \geq v_{i(M)}, \quad v_{i(j)} \in V_i.$$

Здесь индексы (1)..(M) соответствует рейтингу признаков, где признак под индексом (1) соответствует наибольшему вкладу, а (M) – наименьшему.

Для всей выборки определяется средний рейтинг признаков по всем аномальным образцам и составляется множество:

$$G = \{g_1, g_2, \ldots, g_M\},$$

$$g_j = \frac{1}{n}\sum_{i=1}^{n}(j), \quad j=1..M,$$

где (j) – рейтинг j-го признака для i-го образца.

Данный рейтинг и определяет общий вклад каждого признака в результат обнаружения аномалий на всей выборке.

3 Экспериментальная оценка

Задачей поставленных экспериментов является оценка эффективности подхода к объяснимому обнаружению аномалий в ИСИВ с целью повышения надёжности и безопасности производственных процессов. Входными данными являются показатели датчиков ИСИВ, а выходными – набор показателей качества, таких как аккуратность, точность, полнота и F-мера обнаружения аномалий, а также средний рейтинг общего вклада признаков.

3.1 Наборы данных

В качестве экспериментальных наборов данных используются *BATADAL*[6] [38] и *HAI*[7] [39]. Выбор обусловлен намерением сравнить подход к объяснимому обнаружению аномалий на данных ИСИВ с различным количеством признаков и объёмом данных. Указанные наборы часто используются на практике при оценке подходов к обнаружению аномалий в промышленных системах.

[6] *BATADAL* (*BATtle of the Attack Detection Algorithms*). https://www.batadal.net/data.html.
[7] *HAI* (*Hardware-in-the-loop-based Augmented Industrial control system*). https://github.com/icsdataset/hai.

Набор данных *BATADAL* собран на основе показателей индустриальной системы для распределения воды. Хранение и распределение воды по узлам спроса гарантируется семью резервуарами, уровень воды в которых запускает работу одного клапана и одиннадцати насосов, распределённых по пяти насосным станциям. Насосы, клапаны и датчики уровня воды в резервуарах подключены к девяти программируемым логическим контроллерам, которые расположены в непосредственной близости от монитора контроля гидравлических компонентов. Набор *BATADAL* содержит показатели 42 датчиков и актуаторов. Размер нормальной выборки составляет 8 761, а выборки с аномалиями – 4 177 экземпляров.

Наборы данных *HAI* за 2020 (*HAI* 1.0) и 2021 (*HAI* 2.0) годы собраны на испытательном стенде промышленной системы управления, на котором имитируется выработка электроэнергии паровыми турбинами и гидроаккумулирующими электростанциями. Технологический процесс испытательного стенда разделён на четыре процесса: процесс котла (P1), процесс турбины (P2), процесс очистки воды (P3) и аппаратно-программное моделирование (P4). Набор данных *HAI* 1.0 содержит записи 59 точек сбора (показателей датчиков, положения клапанов и т.д.) в выборке 550 800 экземпляров нормальных и 295 000 – аномальных данных. Набор данных *HAI* 2.0 содержит записи 78 точек сбора в выборке 921 603 экземпляров нормальных и 402 005 – аномальных данных. С подробной структурой испытательных стендов и полным наименованием признаков можно ознакомиться в [38, 39]. Далее упоминаются отдельные элементы систем в том виде, в котором они приведены в оригинальных источниках. Технические подробности опущены.

3.2 Параметры моделей реконструкции данных

Размер входного слоя АК (*n_input*) соответствует количеству признаков данных. Для оптимизации гиперпараметров архитектуры АК для каждого набора данных используется байесовский оптимизатор библиотеки *KerasTuner*[8]. Области поиска составляют:

- количество скрытых слоёв (*n_layers*) – 1, 2 или 3;
- количество нейронов на скрытых слоях (*n_units*) – 64, 35, 32, 28 или 25;
- размер скрытого пространства (*encoding_dim*) – 32, 31 или 21;
- функция активации на скрытых слоях (*hidden_afunc*) – *tanh*[9] или *relu*[10];
- функция активации на выходном слое (*output_afunc*)– *tanh* или *relu*.

Параметры обучения на наборе данных *BATADAL*: размер пакета[11] – 32; скорость обучения – 0.001; число эпох – 500. Параметры обучения на наборах данных *HAI*: размер пакета – 64; скорость обучения – 0.001; число эпох – 50. Чтобы ускорить расчёт *SHAP*, используется метод кластеризации *k-средних*. Число кластеров установлено в 100, а параметр *m* принят равным 5.

Описанный подход реализован с использованием языка *Python*.

3.3 Аналоги для сравнения

В качестве подходов к обнаружению аномалий для сравнения полученных результатов выбраны следующие подходы, основанные на методах обучения без учителя:

- *USAD* [19] – подход к обнаружению аномалий, основанный на двух АК в архитектуре *GAN*;
- *VAE* [17] – подход к обнаружению аномалий с использованием вариационного АК;
- *RANSynCoder* [15] – подход к обнаружению аномалий, включающий самонастройку признаков для случайного выбора наборов входных признаков и построения нескольких АК для реконструкции данных временного ряда;
- *InterFusion* [16] – подход к обнаружению аномалий, который основан на использовании иерархического *VAE* с двумя стохастическими скрытыми переменными, каждая из которых изучает низкоразмерные межметрические или временные вложения;
- *FuSAGNet* и *SAE* [20] – подход к обнаружению аномалий, который объединяет разреженный АК и графовую НС, явно моделируя взаимосвязи внутри многомерных временных рядов.

[8] *KerasTuner*. https://keras.io/keras_tuner/.

[9] *Гиперболический тангенс* (англ. *tanh*) – функция активации, которая преобразует входные значения в диапазоне от -1 до 1на основе гиперболического тангенса.

[10] *Линейный выпрямитель* (англ. *Rectified Linear Unit, ReLU*) - это нелинейная функция активации, которая преобразует входное значение в значение от 0 до положительной бесконечности (если входное значение меньше или равно 0, то *ReLU* выдаёт 0, в противном случае – входное значение).

[11] *В глубоком обучении пакет* (или батч) – это подмножество данных, которое используется для обучения модели за один шаг обновления весов.

- *LSTM-FWED* [18] – подход к обнаружению аномалий на основе кодера-декодера *LSTM* с защитой веса признаков от состязательных атак.

3.4 Показатели качества

В качестве функции потерь для всех наборов данных и моделей используется среднеквадратичная ошибка. Показателями качества обнаружения аномалий является аккуратность (*ACC*), точность (*P*), полнота (*R*) и *F*-мера (*F*1):

$$ACC = \frac{TP + TN}{TP + TN + FP + FN},\ P = \frac{TP}{TP + FP},\ R = \frac{TP}{TP + FN},\ F1 = \frac{2TP}{2TP + FP + FN},$$

где *TP* – количество правильно классифицированных аномальных образцов; *TN* – количество правильно классифицированных нормальных образцов; *FP* – количество нормальных образцов, ошибочно классифицированных как аномальные образцы; *FN* – количество аномальных образцов, ошибочно классифицированных как нормальные.

Каждый показатель фокусируется на различных аспектах производительности модели. Аккуратность показывает общую долю правильных предсказаний, но зависит от сбалансированности классов. Точность важна, когда необходимо минимизировать количество ложноположительных результатов, а полнота – когда нужно максимизировать количество правильно идентифицированных положительных случаев. *F*-мера объединяет точность и полноту в одну метрику, что позволяет лучше оценить баланс между ними.

3.5 Результаты

В результате оптимизации гиперпараметров на нормальной выборке получены модели АК, описание архитектуры которых представлено в таблице 1. Целью оптимизации являлась минимизация ошибки реконструкции нормальных данных.

Таблица 1 – Описание архитектуры итоговых моделей автокодировщиков

Модель	***n_input***	***n_layers***	***n_units***	***encoding_dim***	***hidden_afunc***	***output_afunc***
АК-1	32	3	35, 28, 21	21	tanh	tanh
АК-2	58	1	31	31	tanh	tanh
АК-3	78	2	64, 32	32	tanh	tanh

В таблице 2 представлены результаты обнаружения аномалий на выбранных наборах данных. Результаты аналогов для набора данных *BATADAL* взяты из [38], для *HAI* 1.0 из [20], а для *HAI* 2.0 из [40].

Для каждого набора данных выбрано 5 признаков с наибольшей ошибкой реконструкции (в скобках приведён перевод наименований признаков из оригинального описания соответствующих наборов данных):

- *BATADAL* – S_V2 (статус клапана V2), P_J14 (давление сочленения J14), S_PU4 (статус насоса PU4), F_PU6 (поток насоса PU6) и F_PU7 (поток насоса PU7);
- *HAI* 1.0 – P2_SIT01 (текущая частота вращения турбины), P1_PCV02Z (текущее положение клапана PCV02), P1_PCV02D (команда для положения клапана PCV2), P1_FCV03D (команда для положения клапана FCV03) и P1_ FCV03Z (текущее положение клапана FCV03);
- *HAI* 2.0 – P2_VT01 (сигнал фазовой задержки ключевого фазового зонда), P1_PCV02Z (текущее положение клапана PCV02), P2_SIT01, P2_SIT02 (текущая частота вращения турбины) и P1_PCV02D (команда для положения клапана PCV2).

Вектор Шепли для каждого признака можно изобразить в виде графика (см. рисунок 1). Из рисунка видно, как признак с наибольшей ошибкой реконструкции для экземпляра с наибольшей ошибкой влияет на предсказание модели, приближая его к ожидаемому значению. Тёмно-серыми значениями (слева) выделены признаки, которые приближают получен-

ное значение образца к ожидаемому (выделен полужирным), а серыми справа – отдаляющие признаки. Снизу (на графиках - а, б, в) приведены наименования признаков и их значения для выбранного экземпляра. Положительный вклад имеют признаки, которые увеличивают предсказание целевого признака (S_V2 для *BATADAL*, P2_SIT01 для *HAI* 1.0 и P2_VT01 для *HAI* 2.0) и приближают предсказание к ожидаемому значению. Отрицательный вклад имеют признаки, которые уменьшают предсказание модели.

Таблица 2 – Результаты обнаружения аномалий

Набор данных	Модель	*ACC*	*P*	*R*	*F*1
BATADAL	АК-1	0.9725	0.8919	**0.8720**	**0.8818**
	VAE [17]	–	**0.9630**	0.7620	0.8510
	LSTM-FWED [18]	0.9396	0.9228	0.4740	0.6293
HAI 1.0	АК-2	0.9931	**0.9435**	**0.9112**	**0.9271**
	FuSAGNet [20]	–	0.8679	0.7479	0.8034
	SAE [20]	–	0.7839	0.6566	0.7239
	USAD [19].	–	0.0932	0.1335	0.1098
HAI 2.0	АК-3	0.9963	0.8699	**0.9808**	**0.9220**
	RANSynCoder [15]	–	**0.8910**	0.7760	0.8290
	InterFusion [16]	–	0.7440	0.8390	0.7890
	USAD [19].	–	0.7600	0.4800	0.5880

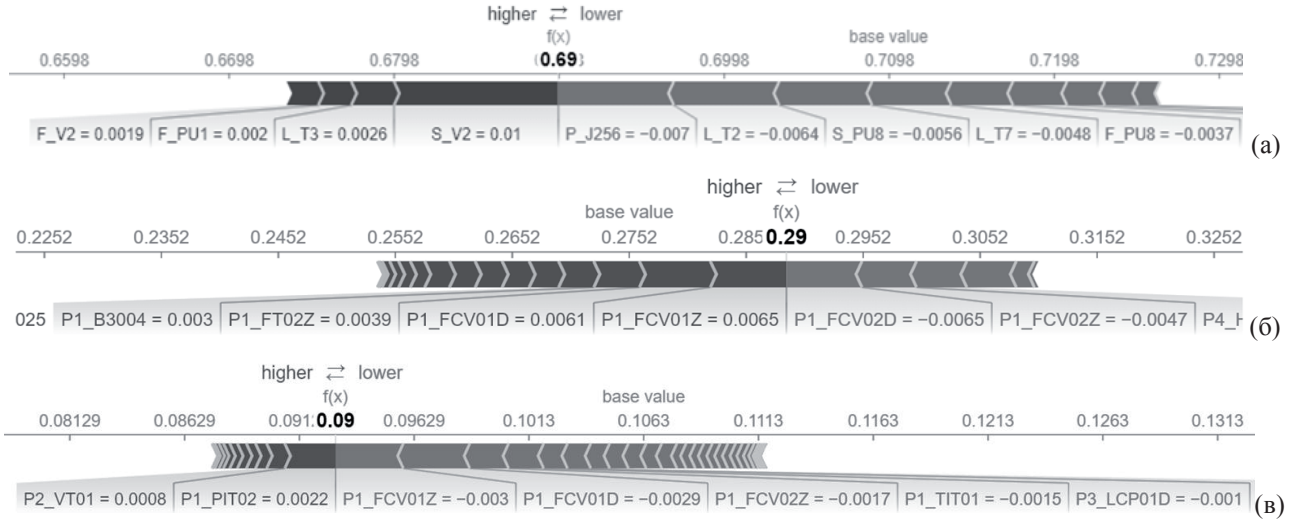
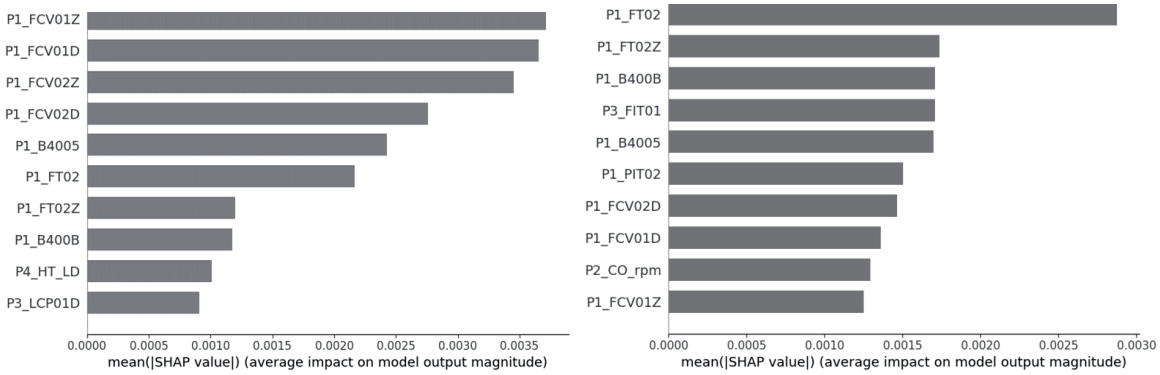


Рисунок 1 – Вектор Шепли для одного образца и признака: (а) S_V2 набора данных *BATADAL*, (б) P2_SIT01 набора данных *HAI* 1.0, (в) P2_VT01 набора данных *HAI* 2.0

Определён вектор средних значений Шепли для каждого образца данных. На рисунке 2 представлен график десяти признаков с наибольшим вкладом для экземпляра с высшей ошибкой реконструкции, т.е. первые 10 значений вектора средних значений Шепли V_i, $i = 234$ для *HAI* 1.0 и $i = 653$ для *HAI* 2.0. Вычислены значения среднего рейтинга каждого признака по всем аномальным экземплярам. На рисунке 3 представлен вклад признаков в обнаружение аномалий на наборе данных *BATADAL*, на рисунке 4 – на наборе данных *HAI* 1.0, на рисунке 5 – на наборе данных *HAI* 2.0.



а) б)

Рисунок 2 – Вклад признаков в ошибку реконструкции для одного образца набора данных (а) *HAI* 1.0 и (б) *HAI* 2.0

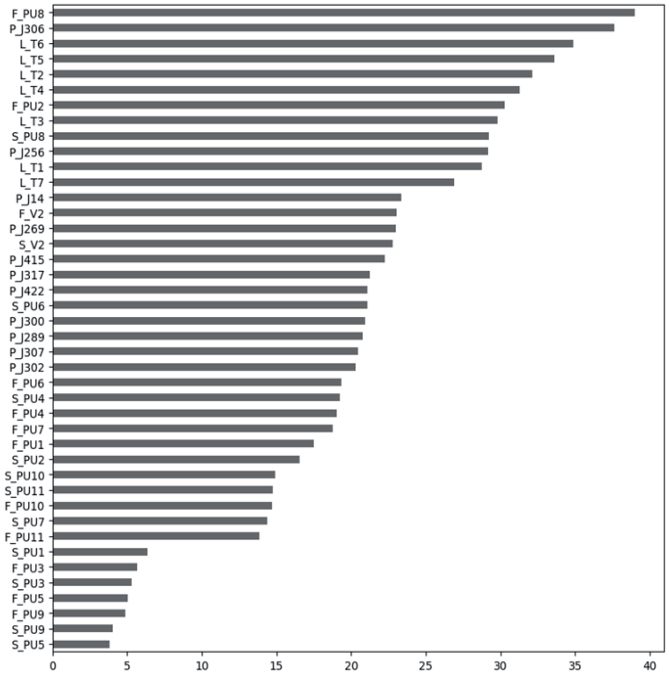


Рисунок 3 – Вклад признаков в ошибку реконструкции на аномальных образцах набора данных *BATADAL*

3.6 Анализ результатов

Предложенный подход позволил получить наибольшую полноту обнаружения аномалий для всех экспериментальных наборов данных 87-98%, а точность обнаружения 87-94%, уступая не более 7% подходам, использованным в [15, 17]. При этом гармоничное среднее по *F*-мере для АК-1, 2 и 3 выше, чем у аналогов (88-93%), что свидетельствует о сбалансированном результате по полноте и точности обнаружения аномалий. Это позволяет минимизировать ложноположительные срабатывания и не пропустить положительные случаи. В задачах обнаружения аномалий значимость ложноположительных и ложноотрицательных срабатываний может значительно различаться.

Например, пропуск аномалии может иметь серьёзные последствия, в то время как ложное срабатывание может быть менее критичным. В производственных процессах пропуск аномалий может привести к поломке оборудования или снижению качества продукции и вызвать значительные финансовые потери. Ложные срабатывания могут привести к ненужным действиям, таким как блокировка процессов, вызов служб безопасности или др. В системах, где критически важно обнаруживать угрозы, часто принимается решение о повышении чувствительности к аномалиям, даже если это может привести к большему количеству ложных срабатываний.

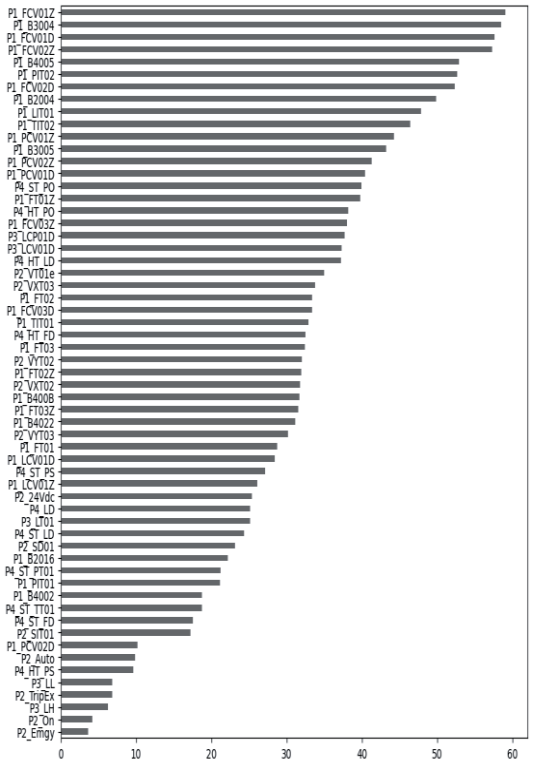


Рисунок 4 – Вклад признаков в ошибку реконструкции на аномальных образцах набора данных HAI 1.0

Рисунок 5 – Вклад признаков в ошибку реконструкции на аномальных образцах набора данных HAI 2.0

Показатель аккуратности редко используется для оценки в задачах обнаружения аномалий. Причиной этому является несбалансированность данных – в большинстве случаев аномалии составляют лишь небольшую часть общего объёма данных. Высокая аккуратность может быть достигнута за счёт предсказания большинства классов без учёта аномалий.

Использование векторов Шепли позволяет продемонстрировать влияние признаков на полученный результат реконструкции. Например, на рисунке 1б можно видеть, что в равной степени (значение Шепли ± 0.0065) на текущую частоту вращения турбины стенда *HAI* влияют текущее положение клапанов FCV01 (P1_FCV01Z) и FCV02 (P1_FCV02D). Значения данных признаков имеют наибольший вклад в ошибку реконструкции экземпляра данных в целом (см. рисунок 2).

Наибольший вклад в обнаружение аномалий на наборе данных *BATADAL* вносят признаки F_PU8 (поток насоса PU8), P_J306 (давление сочленения J306) и L_T6 (уровень воды в баке T6).

Наибольший вклад в обнаружение аномалий на наборе данных *HAI* 1.0 вносят признаки P1_FCV01Z (текущее положение клапана FCV01), P1_B3004 (заданное значение уровня воды), P1_FCV01D (команда для положения клапана FCV01).

Наибольший вклад в обнаружение аномалий на наборе данных *HAI* 2.0 вносят признаки P1_FT02 (измеренный расход воды в баке нагревателя), P1_FT02Z (преобразование расхода воды из P1_FT02) и P1_B4002 (заданное значение температуры на выходе теплообменника). Как и для предыдущей версии набора данных, большое влияние имеют параметры, осуществляющие контроль температуры воды.

Объяснение аномалий помогает исследователям лучше понять поведение интеллектуальных моделей обнаружения, выявить факторы, влияющие на их выводы, и открыть незамеченные ранее закономерности. На практике предложенный подход может способствовать

лучшему пониманию текущих процессов в системах безопасности, помогая обнаруживать угрозы и ошибки в данных.

Заключение

В статье описан подход к обнаружению аномалий при помощи АК и их объяснению на основе метода *SHAP*. АК обучается без учителя, исследуя функцию идентичности данных для реконструкции нормальных экземпляров. В отличие от аналогов, в разработанном подходе предлагается определять вклад признаков для отдельных образцов аномалий и вычислять средний вклад для всей выборки. Понимание того, какие признаки способствуют аномалиям, может помочь в улучшении моделей и алгоритмов для более точного обнаружения аномалий.

Оценка качества предлагаемого подхода проведена на известных наборах данных (*BATADAL*, *HAI* 1.0 и *HAI* 2.0). По итоговой *F*-мере обнаружения достигнут результат в 88-93%, который превосходит рассмотренные аналоги. Показано, как ОИИ может помочь раскрыть причины аномалий в отдельных образцах, и в выборке данных.

Авторский вклад

Котенко И.В. – выбор и постановка задачи исследования; Левшун Д.А., Левшун Д.С. – выбор решений; Левшун Д.А. – программная реализация и проведение экспериментов; Левшун Д.А., Левшун Д.С., Котенко И.В. – обсуждение результатов экспериментов, анализ полученных результатов.

Список источников

[1] ***Levshun D., Chevalier Y., Kotenko I., Chechulin A.*** Design and verification of a mobile robot based on the integrated model of cyber-Physical systems. *Simulation Modelling Practice and Theory*. 2020. Vol.105. P.102151. DOI: 10.1016/j.simpat.2020.102151.

[2] ***Федорченко Е.В., Новикова Е.С., Котенко И.В., Гайфулина Д.А., Тушканова О.Н., Левшун Д.С., Мелешко А.В., Муренин И.Н., Коломеец М.В.*** Система измерения защищенности информации и персональных данных для устройств интернета вещей. *Вопросы кибербезопасности*. 2022. №5. С.28-46. DOI: 10.681/2311-3456-2022-5-28-46.

[3] ***Levshun D., Kotenko I.*** A survey on artificial intelligence techniques for security event correlation: models, challenges, and opportunities. *Artificial Intelligence Review*. 2023. Vol.56(8). P.8547-8590. DOI: 10.1007/s10462-022-10381-4.

[4] ***Котенко И.В., Левшун Д.А.*** Методы интеллектуального анализа системных событий для обнаружения многошаговых кибератак: использование методов машинного обучения. *Искусственный интеллект и принятие решений*. 2023. №3. С. 3-15. DOI: 10.14357/20718594230301.

[5] ***Nwakanma C.I., Ahakonye L.A.C., Njoku J.N., Odirichukwu J.C., Okolie S.A., Uzondu C., Kim D.S.*** Explainable Artificial Intelligence (XAI) for Intrusion Detection and Mitigation in Intelligent Connected Vehicles: A Review. *Applied Sciences*. 2023. Vol.13(3). P.1252. DOI: 10.3390/app13031252.

[6] ***Lundberg S.M., Lee S.I.*** A unified approach to interpreting model predictions. *Advances in neural information processing systems*. 2017. Vol.30. P.1-10. DOI: 10.48550/arXiv.1705.07874.

[7] ***Yang H., Liang S., Ni J., Li H., Shen X.S.*** Secure and efficient k NN classification for industrial Internet of Things. *IEEE Internet of Things Journal*. 2020. Vol.7(11). P.10945-10954. DOI: 10.1109/JIOT.2020.2992349.

[8] ***Hosseinzadeh M., Rahmani A.M., Vo B., Bidaki M., Masdari M., Zangakani M.*** Improving security using SVM-based anomaly detection: issues and challenges. *Soft Computing*. 2021. Vol.25(4). P.3195-3223. DOI: 10.1007/s00500-020-05373-x.

[9] ***Khan M.A., Abuhasel K.A.*** An evolutionary multi-hidden Markov model for intelligent threat sensing in industrial internet of things. *The Journal of Supercomputing*. 2021. Vol.77(6). P.6236-6250. DOI:10.1007/s11227-020-03513-6.

[10] ***Illy P., Kaddoum G., de Araujo-Filho P.F., Kaur K., Garg S.*** A hybrid multistage DNN-based collaborative IDPS for high-risk smart factory networks. *IEEE Transactions on Network and Service Management*. 2022. Vol.19(4). P.4273-4283. DOI: 10.1109/TNSM.2022.3202801.

[11] ***Nandanwar H., Katarya R.*** Deep learning enabled intrusion detection system for Industrial IOT environment. *Expert Systems with Applications*.2024. Vol. 249. P.123808. DOI: 10.1016/j.eswa.2024.123808.
[12] ***Setitra M.A., Fan M., Agbley B.L.Y., Bensalem Z.E.A.*** Optimized MLP-CNN model to enhance detecting DDoS attacks in SDN environment. *Network*. 2023. Vol.3(4). P.538-562. DOI: 10.3390/network3040024.
[13] ***Hasan T., Malik J., Bibi I., Khan W.U., Al-Wesabi F.N., Dev K., Huang G.*** Securing industrial internet of things against botnet attacks using hybrid deep learning approach. *IEEE Transactions on Network Science and Engineering*. 2022. Vol.10(5). P.2952-2963. DOI: 10.1109/TNSE.2022.3168533.
[14] ***Pang G., Shen C., Cao L., Hengel A.V.D.*** Deep learning for anomaly detection: A review. *ACM computing surveys* (CSUR). 2021. Vol.54(2). P.1-38. DOI: 10.1145/3439950.
[15] ***Abdulaal A., Liu Z., Lancewicki T.*** Practical Approach to Asynchronous Multivariate Time Series Anomaly Detection and Localization. *In:. Proc. of the 27th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2021. P.2485-2494. DOI: 10.1145/3447548.3467174.
[16] ***Li Z., Zhao Y., Han J., Su Y., Jiao R., Wen X., Pei D.*** Multivariate Time Series Anomaly Detection and Interpretation using Hierarchical Inter-Metric and Temporal Embedding. *In.: Proc. of the 27th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2021. P.3220-3230. DOI: 10.1145/3447548.3467075.
[17] ***Kravchik M., Shabtai A.*** Efficient cyber attack detection in industrial control systems using lightweight neural networks and PCA. *IEEE transactions on dependable and secure computing*. 2021. Vol.19(4). P.2179-2197. DOI: 10.1109/TDSC.2021.3050101.
[18] ***Liu Y., Xu L., Yang, S., Zhao D., Li X.*** Adversarial sample attacks and defenses based on LSTM-ED in industrial control systems // Computers & Security. 2024. Vol. 140. P.103750. DOI: 10.1016/j.cose.2024.103750.
[19] ***Audibert J., Michiardi P., Guyard F., Marti S, Zuluaga M.A.*** USAD: Unsupervised Anomaly Detection on Multivariate Time Series. *In.: Proc. of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2020. P. 3395-3404. DOI: 10.1145/3394486.3403392.
[20] ***Han S., Woo S.S.*** Learning sparse latent graph representations for anomaly detection in multivariate time series. *In.: Proc. of the 28th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2022. P.2977-2986. DOI: 10.1145/3534678.3539117.
[21] ***Utkin L., Konstantinov A.*** An Extension of the Neural Additive Model for Uncertainty Explanation of Machine Learning Survival Models // Cyber-Physical Systems: Intelligent Models and Algorithms. Cham : Springer International Publishing, 2022. P.3-13. DOI: 10.1007/978-3-030-95116-0_1.
[22] ***Chen Q., Pan G., Chen W., Wu P.*** A novel explainable deep belief network framework and its application for feature importance analysis. *IEEE Sensors Journal*. 2021. Vol. 21(22). P.25001-25009. DOI: 10.1109/JSEN.2021.3084846.
[23] ***Zolanvari M., Yang Z., Khan K., Jain R., Meskin N.*** TRUST XAI: Model-agnostic explanations for ai with a case study on iiot security. *IEEE Internet of Things Journal*. 2023. Vol.10(4). P.2967-2978. DOI: 10.1109/JIOT.2021.3122019.
[24] ***Ribeiro M.T., Singh S., Guestrin C.*** "Why should i trust you?" Explaining the predictions of any classifier. *In.: Proc. of the 22nd ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2016. P.1135-1144. DOI: 10.1145/2939672.2939778.
[25] ***Liu M., Shi J., Cao K., Zhu J., Liu S.*** Analyzing the training processes of deep generative models. *IEEE transactions on visualization and computer graphics*. 2017. Vol.24(1). P.77-87. DOI: 10.1109/TVCG.2017.2744938.
[26] ***Wang P.Y., Galhotra S., Pradhan R., Salimi B.*** Demonstration of generating explanations for black-box algorithms using Lewis. *In.: Proc. of the VLDB Endowment*. 2021. Vol.14(12). P.2787-2790. DOI: 10.14778/3476311.3476345.
[27] ***Moradi M., Samwald M.*** Post-hoc explanation of black-box classifiers using confident item sets. *Expert Systems with Applications*. 2021. Vol.165. P.113941. DOI: 10.1016/j.eswa.2020.113941.
[28] ***Nourani M., Roy C., Block J. E., Honeycutt D. R., Rahman T., Ragan E., Gogate V.*** Anchoring bias affects mental model formation and user reliance in explainable ai systems. *In.: Proc. of the 26th International Conference on Intelligent User Interfaces*. 2021. P.340-350. DOI: 10.1145/3397481.3450639.
[29] ***Abou El Houda Z., Brik B., Senouci S.M.*** A novel iot-based explainable deep learning framework for intrusion detection systems. *IEEE Internet of Things Magazine*. 2022. Vol.5(2). P.20-23. DOI: 10.1109/IOTM.005.2200028.
[30] ***Kopp M., Pevný T., Holeňa M.*** Anomaly explanation with random forests. *Expert Systems with Applications*. 2020. Vol.149. P.113187. DOI: 10.1016/j.eswa.2020.113187.
[31] ***Antwarg L., Miller R.M., Shapira B., Rokach L.*** Explaining anomalies detected by autoencoders using Shapley Additive Explanations // Expert systems with applications. 2021. Vol. 186. P.115736. DOI: 10.1016/j.eswa.2020.113187.

[32] ***Nguyen Q.P., Lim K.W., Divakaran D.M., Low K.H., Chan M.C.*** Gee: A gradient-based explainable variational autoencoder for network anomaly detection. *In.: Proc of the 2019 IEEE Conference on Communications and Network Security* (CNS). 2019. P.91-99. DOI: 10.1109/CNS.2019.8802833.

[33] ***Takeishi N.*** Shapley values of reconstruction errors of PCA for explaining anomaly detection. *In.: Proc of the 2019 international conference on data mining workshops* (ICDMW). 2019. P.793-798. DOI: 10.1109/ICDMW.2019.00117.

[34] ***Roshan K., Zafar A.*** Utilizing XAI technique to improve autoencoder based model for computer network anomaly detection with shapley additive explanation (SHAP). *International Journal of Computer Networks & Communications* (IJCNC). 2021. Vol.13(6). P.1-20. DOI: 10.5121/ijcnc.2021.13607.

[35] ***Huong T.T., Bac T.P., Ha K.N., Hoang N.V., Hoang N.X., Hung N.T., Tran K.P.*** Federated learning-based explainable anomaly detection for industrial control systems. *IEEE Access*. 2022. Vol.10. P.53854-53872. DOI: 10.1109/ACCESS.2022.3173288.

[36] ***Mathuros K., Venugopalan S., Adepu S.*** WaXAI: Explainable Anomaly Detection in Industrial Control Systems and Water Systems. *In.: Proceedings of the 10th ACM Cyber-Physical System Security Workshop*. 2024. P.3-15. DOI: 10.1145/3626205.3659147.

[37] ***Snoek J., Larochelle H., Adams R. P.*** Practical bayesian optimization of machine learning algorithms. *Advances in neural information processing systems.* 2012. Vol.25. P.1-9. DOI: 10.48550/arXiv.1206.2944.

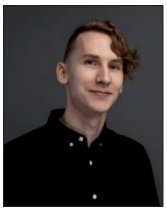
[38] ***Taormina R., Galelli S., Tippenhauer N. O., Salomons E., Ostfeld A., Eliades D. G***. Battle of the attack detection algorithms: Disclosing cyber attacks on water distribution networks. *Journal of Water Resources Planning and Management*. 2018. Vol.144(8). P.04018048. DOI: 10.1061/(ASCE)WR.1943-5452.0000969.

[39] ***Shin H.K., Lee W., Yun J.H., Min B.G.*** Two ICS security datasets and anomaly detection contest on the HIL-based augmented ICS testbed. *In.: Proceedings of the 14th Cyber Security Experimentation and Test Workshop.* 2021. P.36-40. DOI: 10.1145/3474718.3474719.

[40] ***Kim B., Alawami M.A., Kim E., Oh S., Park J., Kim H.*** A comparative study of time series anomaly detection models for industrial control systems. *Sensors*. 2023. Vol.23(3). P.1310. DOI: 10.3390/s23031310.

Сведения об авторах

Левшун Диана Альбертовна, 1995 г. рождения. С отличием окончила бакалавриат Оренбургского государственного университета (2017) и магистратуру Университета ИТМО в 2019 г. Младший научный сотрудник лаборатории проблем компьютерной безопасности СПб ФИЦ РАН. В списке научных трудов около 70 работ в области информационной безопасности и искусственного интеллекта. ORCID: 0000-0002-5266-8649; Author ID (РИНЦ): 968755; Author ID (Scopus): 58114512500; Researcher ID (WoS): ABG-9837-2020. *gaifulina@comsec.spb.ru*.



Левшун Дмитрий Сергеевич, 1993 г. рождения. Выпускник СПбГЭТУ «ЛЭТИ» (2017), к.т.н. (2021). Старший научный сотрудник лаборатории проблем компьютерной безопасности СПб ФИЦ РАН. Доцент СПбГУТ и Европейского университета в Санкт-Петербурге. В списке научных трудов более 100 работ в области информационной безопасности, проектирования защищенных систем, Интернета вещей, искусственного интеллекта, моделирования атак и атакующих. ORCID: 0000-0003-1898-6624; Author ID (РИНЦ): 840344; Author ID (Scopus): 57189306576; Researcher ID (WoS): C-1566-2018. *levshun@comsec.spb.ru.*

Котенко Игорь Витальевич, 1961 г. рождения. С отличием окончил Военный инженерный Краснознамённый институт им. А.Ф. Можайского в 1983 г. и Военную академию связи в 1987 г., д.т.н. (1999), профессор (2021), заслуженный деятель науки Российской федерации (2023). Главный научный сотрудник и руководитель Лаборатории проблем компьютерной безопасности СПб ФИЦ РАН, профессор Университета ИТМО, СПбГУТ, УрФУ, Харбинского политехнического университета (КНР) и Хэйлунцзянского университета (КНР). В списке научных трудов более 800 работ в области безопасности компьютерных сетей, искусственного интеллекта, телекоммуникационных систем, включая 25 монографий, и более 100 патентов на изобретения и зарегистрированных программ. ORCID: 0000-0001-6859-7120; Author ID (РИНЦ): 110102; Author ID (Scopus): 15925268000. *ivkote@comsec.spb.ru.* ✉.

Поступила в редакцию 22.10.2024, после рецензирования 2.12.2024. Принята к публикации 14.01.2025.

Scientific article DOI: 10.18287/2223-9537-2025-15-1-96-113

Detecting and explaining anomalies in industrial Internet of things systems using an autoencoder

© 2025, D.A. Levshun, D.S. Levshun, I.V. Kotenko ✉

St. Petersburg Federal Research Center of the Russian Academy of Sciences, St. Petersburg, Russia

Abstract

In industrial Internet of Things (IoT) systems, explaining anomalies plays a crucial role in identifying bottlenecks and optimizing processes. This paper proposes an approach to anomaly detection using an autoencoder and its explanation based on the SHAP method. The purpose of the anomaly explanation is to provide a set of data features in industrial IoT systems that most significantly influence anomaly detection. The novelty of this approach lies in its ability to quantify the contribution of individual features for specific data samples and to calculate an average contribution across the dataset, providing a feature importance ranking. The proposed approach is tested on Industrial IoT datasets with varying feature counts and data volumes. The anomaly detection achieves an F-measure of 88-93%, outperforming the comparable methods discussed. The study demonstrates how explainable artificial intelligence can identify the causes of anomalies in both individual samples and datasets as a whole. The theoretical importance of the proposed approach lies in its ability to shed light on the workings of intelligent detection models, enabling the identification of factors influencing their outcomes and uncovering previously unnoticed patterns. In practice, this method enhances security system operators' understanding of ongoing processes, aiding in threat identification and error detection within data.

Keywords: *information security, anomaly detection, industrial IoT systems, autoencoder, explainable artificial intelligence.*

For citation: *Levshun DA, Levshun DS, Kotenko IV.* Detecting and explaining anomalies in industrial Internet of things systems using an autoencoder [In Russian]. *Ontology of designing*. 2025; 15(1): 96-113. DOI:10.18287/2223-9537-2025-15-1-96-113.

Financial Support: The study was supported by the grant of the St. Petersburg Science Foundation No. 23-РБ-01-09.

Conflict of interest: The authors declare no conflict of interest.

List of figures and tables

References

[1] ***Levshun D, Chevalier Y, Kotenko I, Chechulin A*** Design and verification of a mobile robot based on the integrated model of cyber-Physical systems. *Simulation Modelling Practice and Theory*. 2020; 105: 102151. DOI: 10.1016/j.simpat.2020.102151.

[2] ***Fedorchenko EV, Novikova ES, Kotenko IV, Gaifulina DA, Tushkanova ON, Levshun DS, Meleshko AV, Murenin IN, Kolomeec MV.*** The security and privacy measuring system for the internet of things devices [In Russian]. *Voprosy kiberbezopasnosti*. 2022; 5: 28-46. DOI: 10.681/2311-3456-2022-5-28-46.

[3] ***Levshun D, Kotenko I.*** A survey on artificial intelligence techniques for security event correlation: models, challenges, and opportunities. *Artificial Intelligence Review*. 2023; 56(8): 8547-8590. DOI: 10.1007/s10462-022-10381-4.

[4] ***Kotenko IV, Levshun DA.*** Methods of intelligent system event analysis for multistep cyber-attack detection: using machine learning methods [In Russian]. *Iskusstvenniy Intellekt i Prinyatie Resheniy*. 2023: 3: 3-15. DOI: 10.14357/20718594230301.

[5] ***Nwakanma CI, Ahakonye LAC, Njoku JN, Odirichukwu JC, Okolie SA, Uzondu C, Kim DS.*** Explainable Artificial Intelligence (XAI) for Intrusion Detection and Mitigation in Intelligent Connected Vehicles: A Review. *Applied Sciences*. 2023; 13(3): 1252. DOI: 10.3390/app13031252.

[6] ***Lundberg SM, Lee SI.*** A unified approach to interpreting model predictions. Advances in neural information processing systems. 2017; 30: 1-10. DOI: 10.48550/arXiv.1705.07874.

[7] ***Yang H, Liang S, Ni J, Li H, Shen XS.*** Secure and efficient k-NN classification for industrial Internet of Things. *IEEE Internet of Things Journal*. 2020; 7(11): 10945-10954. DOI: 10.1109/JIOT.2020.2992349.

[8] ***Hosseinzadeh M, Rahmani A M, Vo B, Bidaki M, Masdari M, Zangakani M*** Improving security using SVM-based anomaly detection: issues and challenges. *Soft Computing* 2021; 25(4): 3195-3223. DOI: 10.1007/s00500-020-05373-x.

[9] ***Khan M A, Abuhasel K A.*** An evolutionary multi-hidden Markov model for intelligent threat sensing in industrial internet of things. *The Journal of Supercomputing*. 2021; 77(6): 6236-6250. DOI:10.1007/s11227-020-03513-6.

[10] ***Illy P, Kaddoum G, de Araujo-Filho, P F, Kaur K, Garg S.*** A hybrid multistage DNN-based collaborative IDPS for high-risk smart factory networks. *IEEE Transactions on Network and Service Management*. 2022; 19(4): 4273-4283. DOI: 10.1109/TNSM.2022.3202801.

[11] ***Nandanwar H, Katarya R.*** Deep learning enabled intrusion detection system for Industrial IOT environment. *Expert Systems with Applications*. 2024; 249: 123808. DOI: 10.1016/j.eswa.2024.123808.

[12] ***Setitra MA, Fan M, Agbley BLY, Bensalem ZEA.*** Optimized MLP-CNN model to enhance detecting DDoS attacks in SDN environment. *Network*. 2023; 3(4): 538-562. DOI: 10.3390/network3040024.

[13] ***Hasan T, Malik J, Bibi I, Khan W U, Al-Wesabi F N, Dev K, Huang G.*** Securing industrial internet of things against botnet attacks using hybrid deep learning approach. *IEEE Transactions on Network Science and Engineering*. 2022; 10(5): 2952-2963. DOI: 10.1109/TNSE.2022.3168533.

[14] ***Pang G, Shen C, Cao L, Hengel AVD.*** Deep learning for anomaly detection: A review. *ACM computing surveys* (CSUR). 2021; 54(2): 1-38. DOI: 10.1145/3439950.

[15] ***Abdulaal A, Liu Z, Lancewicki T.*** Practical Approach to Asynchronous Multivariate Time Series Anomaly Detection and Localization. In: *Proc of the 27th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2021; 2485-2494. DOI: 10.1145/3447548.3467174.

[16] ***Li Z, Zhao Y, Han J, Su Y, Jiao R, Wen X, Pei D.*** Multivariate Time Series Anomaly Detection and Interpretation using Hierarchical Inter-Metric and Temporal Embedding. In: *Proc of the 27th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2021: 3220-3230. DOI: 10.1145/3447548.3467075.

[17] ***Kravchik M, Shabtai A.*** Efficient cyber attack detection in industrial control systems using lightweight neural networks and PCA. *IEEE transactions on dependable and secure computing*. 2021; 19(4): 2179-2197. DOI: 10.1109/TDSC.2021.3050101.

[18] ***Liu Y, Xu L, Yang, S, Zhao D, Li X.*** Adversarial sample attacks and defenses based on LSTM-ED in industrial control systems. *Computers & Security*. 2024; 140: 103750. DOI: 10.1016/j.cose.2024.103750.

[19] ***Audibert J, Michiardi P, Guyard F, Marti S, Zuluaga MA.*** USAD: Unsupervised Anomaly Detection on Multivariate Time Series. In: Proc of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining; 2020 P 3395-3404. DOI: 10.1145/3394486.3403392.

[20] ***Han S, Woo SS.*** Learning sparse latent graph representations for anomaly detection in multivariate time series. In: Proc of the 28th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2022; 2977-2986. DOI: 10.1145/3534678.3539117.

[21] ***Utkin L, Konstantinov A.*** An Extension of the Neural Additive Model for Uncertainty Explanation of Machine Learning Survival Models. Cyber-Physical Systems: Intelligent Models and Algorithms Cham : Springer International Publishing. 2022; 3-13. DOI: 10.1007/978-3-030-95116-0 1.

[22] ***Chen Q, Pan G, Chen W, Wu P.*** A novel explainable deep belief network framework and its application for feature importance analysis. IEEE Sensors Journal. 2021; 21(22): 25001-25009. DOI: 10.1109/JSEN.2021.3084846.

[23] ***Zolanvari M, Yang Z, Khan K, Jain R, Meskin N.*** TRUST XAI: Model-agnostic explanations for ai with a case study on iiot security. *IEEE Internet of Things Journal*. 2023; 10(4); 2967-2978. DOI: 10.1109/JIOT.2021.3122019.

[24] ***Ribeiro M T, Singh S, Guestrin C.*** "Why should i trust you?" Explaining the predictions of any classifier. In: Proc of the 22nd ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2016; 1135-1144. DOI: 10.1145/2939672.2939778.

[25] ***Liu M, Shi J, Cao K, Zhu J, Liu S.*** Analyzing the training processes of deep generative models. IEEE transactions on visualization and computer graphics. 2017; 24(1): 77-87. DOI: 10.1109/TVCG.2017.2744938.

[26] ***Wang P Y, Galhotra S, Pradhan R, Salimi B.*** Demonstration of generating explanations for black-box algorithms using Lewis. In: Proc of the VLDB Endowment. 2021; 14(12): 2787-2790. DOI: 10.14778/3476311.3476345.

[27] ***Moradi M, Samwald M.*** Post-hoc explanation of black-box classifiers using confident itemsets. Expert Systems with Applications. 2021; 165: 113941. DOI: 10.1016/j.eswa.2020.113941.

[28] ***Nourani M, Roy C, Block JE, Honeycutt DR, Rahman T, Ragan E, Gogate V.*** Anchoring bias affects mental model formation and user reliance in explainable ai systems. In: Proc of the 26th International Conference on Intelligent User Interfaces. 2021; 340-350. DOI: 10.1145/3397481.3450639.

[29] ***Abou El Houda Z, Brik B, Senouci SM.*** A novel iot-based explainable deep learning framework for intrusion detection systems. IEEE Internet of Things Magazine. 2022; 5(2): 20-23. DOI: 10.1109/IOTM.005.2200028.

[30] ***Kopp M, Pevný T, Holeňa M.*** Anomaly explanation with random forests. Expert Systems with Applications 2020; 149: 113187. DOI: 10.1016/j.eswa.2020.113187.

[31] ***Antwarg L, Miller R M, Shapira B, Rokach L.*** Explaining anomalies detected by autoencoders using Shapley Additive Explanations. Expert systems with applications. 2021; 186: 115736. DOI: 10.1016/j.eswa.2020.113187.

[32] ***Nguyen QP, Lim KW, Divakaran DM, Low KH, Chan MC.*** Gee: A gradient-based explainable variational autoencoder for network anomaly detection. In: Proc of the 2019 IEEE Conference on Communications and Network Security (CNS). 2019; 91-99. DOI: 10.1109/CNS.2019.8802833.
[33] ***Takeishi N.*** Shapley values of reconstruction errors of PCA for explaining anomaly detection. In: Proc of the 2019 international conference on data mining workshops (ICDMW). 2019; 793-798. DOI: 10.1109/ICDMW.2019.00117.
[34] ***Roshan K, Zafar A.*** Utilizing XAI technique to improve autoencoder based model for computer network anomaly detection with shapley additive explanation (SHAP). International Journal of Computer Networks & Communications (IJCNC). 2021; 13(6): 1-20. DOI: 10.5121/ijcnc.2021.13607.
[35] ***Huong TT, Bac TP, Ha KN, Hoang NV, Hoang NX, Hung NT, Tran KP.*** Federated learning-based explainable anomaly detection for industrial control systems. IEEE Access. 2022; 10: 53854-53872. DOI: 10.1109/ACCESS.2022.3173288.
[36] ***Mathuros K, Venugopalan S, Adepu S.*** WaXAI: Explainable Anomaly Detection in Industrial Control Systems and Water Systems. In: Proceedings of the 10th ACM Cyber-Physical System Security Workshop. 2024; 3-15. DOI: 10.1145/3626205.3659147.
[37] ***Snoek J., Larochelle H., Adams R. P.*** Practical bayesian optimization of machine learning algorithms. Advances in neural information processing systems. 2012; 25: 1-9.
[38] ***Taormina R, Galelli S, Tippenhauer NO, Salomons E, Ostfeld A, Eliades DG.*** Battle of the attack detection algorithms: Disclosing cyber attacks on water distribution networks. Journal of Water Resources Planning and Management. 2018; 144(8): 04018048. DOI: 10.1061/(ASCE)WR.1943-5452.0000969.
[39] ***Shin HK, Lee W, Yun JH, Min BG.*** Two ICS security datasets and anomaly detection contest on the HIL-based augmented ICS testbed. In: *Proc. of the 14th Cyber Security Experimentation and Test Workshop*. 2021; 36-40. DOI: 10.1145/3474718.3474719.
[40] ***Kim B, Alawami MA, Kim E, Oh S, Park J, Kim H.*** A comparative study of time series anomaly detection models for industrial control systems. Sensors. 2023; 23(3): 1310. DOI: 10.3390/s23031310.

About the authors

Diana Albertovna Levshun (b. 1995) graduated with a bachelor's degree from Orenburg State University in 2017 and received a master's degree in 2019 from ITMO University. She is a Junior Researcher at the Laboratory of Computer Security Problems at the St. Petersburg Federal Research Center of the Russian Academy of Sciences. Her list of scientific publications includes about 70 works in the field of information security and artificial intelligence. ORCID: 0000-0002-5266-8649; Author ID (RSCI): 968755; Author ID (Scopus): 58114512500; Researcher ID (WoS): ABG-9837-2020. *gaifulina@comsec.spb.ru*.

Dmitry Sergeevich Levshun (b 1993) graduated from Saint Petersburg Electrotechnical University "LETI" named after V.I. Ulyanov (Lenin) in 2017, Ph.D. (2021). He is a Senior Researcher at the Laboratory of Computer Security Problems at the St. Petersburg Federal Research Center of the Russian Academy of Sciences. He is an Associate Professor at the Saint Petersburg State University of Telecommunications named after prof. M.A. Bonch-Bruevich and at the European University at St. Petersburg. His list of scientific publications includes more than 100 works in the field of information security, security by design, Internet of Things, artificial intelligence, modelling of attacks and attackers. ORCID: 0000-0003-1898-6624; Author ID (RSCI): 840344; Author ID (Scopus): 57189306576; Researcher ID (WoS): C-1566-2018. *levshun@comsec.spb.ru.*

Igor Vitalievich Kotenko (b. 1961) graduated with honors from the St. Petersburg Academy of Space Engineering in 1983 and St. Petersburg Signal Academy in 1987, D. Sc. Eng. (1999), professor (2021), Honored Scientist of the Russian Federation (2023). Chief Researcher and Head of the Laboratory of Computer Security Problems at St. Petersburg Federal Research Center of the Russian Academy of Sciences, professor at ITMO University, Saint Petersburg State University of Telecommunications, Ural Federal University, Harbin Institute of Technology (China) and Heilongjiang University (China). He is a co-author of more than 800 publications in the field of computer network security, artificial intelligence, telecommunication systems, including 25 monographs, and more than 100 patents for inventions and registered programs. ORCID: 0000-0001-6859-7120; Author ID (RSCI): 110102; Author ID (Scopus): 15925268000. *ivkote@comsec.spb.ru.* ✉.

Received October 22, 2024, Revised December 2, 2024 Accepted January 14, 2025