ПОДХОД К ОБЪЯСНИМОМУ ОБНАРУЖЕНИЮ АНОМАЛИЙ В ПОТОКЕ ДАННЫХ ОТ ТЕХНОЛОГИЧЕСКИХ ПРОЦЕССОВ

Новикова Е. С.[1], Бухтияров М. А.[2], Котенко И. В.[3], Саенко И. Б.[4], Федорченко Е. В.[5]

DOI: 10.21681/2311-3456-2025-4-142-151

Цель исследования: *разработка подхода к выявлению аномалий в данных технологических процессов на основе объяснимого машинного обучения в целях дальнейшего выбора контрмер с учетом возможных источников аномалий.*

Методы исследования: *статистический анализ, методы машинного обучения, методы генерации объяснений к прогнозам модели машинного обучения.*

Полученные результаты: *предложен подход к объяснимому обнаружению аномалий в потоке данных от технологических процессов, и представлены его основные этапы, в основе которых лежит преобразование входного вектора данных в матрицу и выявление аномалий с помощью сверточной нейронной сети; разработана методика трансформации вектора данных в матрицу, и оценено влияние алгоритма преобразования данных на эффективность решения задачи выявления аномалий; разработана методика тестирования точности генерируемых объяснений и выполнена экспериментальная оценка методов SHAP, Grad-CAM и Guided Grad-CAM.*

Научная новизна: *предложенный подход к выявлению аномалий в данных технологического процесса отличается от существующих использованием разработанной методики преобразования вектора входных данных в матрицу, что позволяет применить сверточную нейронную сеть в качестве аналитической модели выявления аномалий и методы генерации объяснений, разработанные специально для нейронных сетей данной архитектуры.*

Вклад: *Новикова Е. С. – разработка методики преобразования входного потока данных; Бухтияров М. А. – экспериментальное исследование предложенного подхода; Котенко И. В. – разработка общего подхода к объяснимому обнаружению аномалий в рамках концепции динамического оценивания защищенности информационных систем в условиях неопределенности исходных данных; Котенко И. В., Саенко И. Б. и Федорченко Е. В. – анализ современных исследований по выявлению аномалий в технологических процессах и формированию объяснений к прогнозам моделей машинного обучения.*

Ключевые слова: *обнаружение кибератак и аномалий, промышленные киберфизические системы, генерация аномалий, оценка точности объяснений.*

Введение

Цифровая трансформация производственных систем связана с внедрением технологий Интернета вещей, которые позволяют усовершенствовать производственные процессы, повысить эффективность их управления, осуществлять мониторинг состояния оборудования [1]. Однако интеграция сетевых технологий, обеспечивающих в том числе удаленное подключение к корпоративным информационным системам, приводит к тому, что промышленные системы управления сталкиваются с повышенными рисками информационной безопасности [2]. Обеспечение безопасности таких систем является критически важной задачей, поскольку последствия реализаций информационных угроз могут нанести серьезный экономический и экологический ущерб. Например, нарушение процессов водоочистки или водоподготовки воды могут привести не только к значительным сбоям в работе этих систем, но и к загрязнению источников водоснабжения и потенциальной опасности для здоровья [3].

Для своевременного обнаружения аномалий в технологических процессах предложены разнообразные методы как на основе статистического анализа данных, так и на основе машинного обучения, в том числе глубокого обучения [4, 5]. Методы на основе глубокого обучения показали высокую эффективность решения данной задачи, однако их применение значительно усложняет анализ первопричин аномалий несмотря на то, что эта задача важна для промышленных киберфизических систем (КФС) [6]. Определение источника аномалий входит в процедуры оценки рисков, в частности, на основе этих

1 Новикова Евгения Сергеевна, кандидат технических наук, старший научный сотрудник, ФГБУН «Санкт-Петербургский Федеральный исследовательский центр Российской академии наук» (СПб ФИЦ РАН), г. Санкт-Петербург, Россия. E-mail: novikova@comsec.spb.ru

2 Бухтияров Марат Андреевич, программист, ФГБУН «Санкт-Петербургский Федеральный исследовательский центр Российской академии наук» (СПб ФИЦ РАН), г. Санкт-Петербург, Россия. E-mail: buhtiarov.marat@gmail.com

3 Котенко Игорь Витальевич, заслуженный деятель науки РФ, доктор технических наук, профессор, главный научный сотрудник и руководитель лаборатории проблем компьютерной безопасности, ФГБУН «Санкт Петербургский Федеральный исследовательский центр Российской академии наук» (СПб ФИЦ РАН), г. Санкт-Петербург, Россия. E-mail: ivkote@comsec.spb.ru

4 Саенко Игорь Борисович, доктор технических наук, профессор, главный научный сотрудник, ФГБУН «Санкт Петербургский Федеральный исследовательский центр Российской академии наук» (СПб ФИЦ РАН), г. Санкт-Петербург, Россия. E-mail: ibsaen@comsec.spb.ru

5 Федорченко Елена Владимировна, кандидат технических наук, старший научный сотрудник, ФГБУН «Санкт Петербургский Федеральный исследовательский центр Российской академии наук» (СПб ФИЦ РАН), г. Санкт-Петербург, Россия. E-mail: doynikova@comsec.spb.ru

данных вычисляются оценки киберрисков для активов организации и формируются возможные контрмеры [7, 8].

В настоящей работе предлагается подход к объяснимому выявлению аномалий, который включает этапы предобработки входных данных, выявления аномалий методами машинного обучения и генерации объяснений к прогнозам обученной модели. Отличительной особенностью подхода является преобразование входного вектора данных от технологических процессов в матрицу, что позволяет применить сверточные нейронные сети и методы генерации объяснений, разработанные для нейронных сетей с данной архитектурой. В работе исследуется точность различных методов генерации объяснений, в частности рассмотрены метод SHAP, который не зависит от архитектуры нейронной сети, и методы Grad-CAM (Gradient-weighted Class Activation Mapping, градиентно-взвешенное отображение активации класса) и Guided Grad-CAM (управляемое градиентно-взвешенное отображение активации класса), разработанные специально для сверточных нейронных сетей. Таким образом, основным вкладом авторов являются: разработка общего подхода к объяснимому выявлению аномалий в потоке данных от технологических процессов; методика выявления аномалий на основе преобразования входного вектора данных в матрицу и сверточной нейронной сети; сравнительный анализ методов генерации объяснений в задачах выявления аномалий в многомерных временных рядах.

Работа построена следующим образом. В разделе 2 обсуждаются исследования в области обнаружения аномалий в технологических процессах и методы генерации объяснений. В разделе 3 представлен разработанный подход к объяснимому обнаружению аномалий, описаны основные его этапы. В разделе 4 рассмотрен сценарий эксперимента, и обсуждаются полученные результаты эффективности обнаружения аномалий и точности генерируемых объяснений. В заключении представлены основные результаты и определяются дальнейшие направления работ.

Анализ релевантных работ

Основным типом данных от КФС являются многомерные временные ряды. При их анализе необходимо учитывать взаимосвязи между различными атрибутами, которые могут быть нелинейными и динамически развивающимися во времени, что обеспечивает выявление аномалий традиционными методами. В последнее время для выявления аномалий предложены методы на основе глубоких нейронных сетей [9, 10].

В частности, в [11] для выявления аномалий во временных рядах представлено решение OmniAnomaly, в основе которого лежит стохастическая рекуррентная нейронная сеть и вариационный автокодировщик для извлечения временных зависимостей между атрибутами. В [12] предложен подход, основанный на применении сверточной нейронной сети и двух автокодировщиков со слоями долгой короткосрочной памяти. Автокодировщики используются для обнаружения аномалий и редких событий путем выявления краткосрочных и долгосрочных отклонений фактических значений датчиков от прогнозируемых значений. Похожее решение представлено в [13], однако сверточные слои нейронной сети здесь дополнены механизмом внимания, что позволяет сфокусировать акцент сети на наиболее важных извлекаемых признаках.

Джао и др. [14] адаптировали генеративные состязательные сети для решения задачи обнаружения аномалий в условиях несбалансированных наборов данных, причем при генерации синтетических данных реализован принцип полного ассоциативного отображения, то есть нормальные данные используются для генерации аномальных данных и наоборот.

Применение методов глубокого обучения для обнаружения аномалий во временных рядах усложняет определение источника аномалий. В зависимости от подхода и целей существует две основные группы методов, которые могут объяснить предсказания модели машинного обучения: методы, учитывающие специфику архитектуры модели, и методы, не зависящие от модели.

Методы, учитывающие специфику конкретной модели, учитывают встроенные свойства модели для формирования объяснимости. К таким методам относятся методы на основе построения карт значимости (Class Activation Maps, CAM, карты активации классов), разработанных для сверточных нейронных сетей и рассчитываемых на основе оценки градиентов; методы послойного распространения релевантности (Layer-wise Relevance Propagation, LRP), применяемые для анализа прогнозов рекуррентных и сверточных нейронных сетей; методы на основе анализа механизмов внимания.

Методы, не зависящие от модели, позволяют объяснить предсказания моделей машинного обучения, не опираясь на специфические свойства этих моделей, и могут применяться к любой модели, независимо от используемых алгоритмов обучения. Методы этой группы, как правило, работают после обучения основной модели (post-hoc методы) и не влияют на процесс обучения и генерацию предсказаний. К таким методам относятся метод аддитивных объяснений на основе вектора Шэпли (SHAP) [15] и метод модельно-независимых локальных объяснений (LIME) [16], которые широко используются

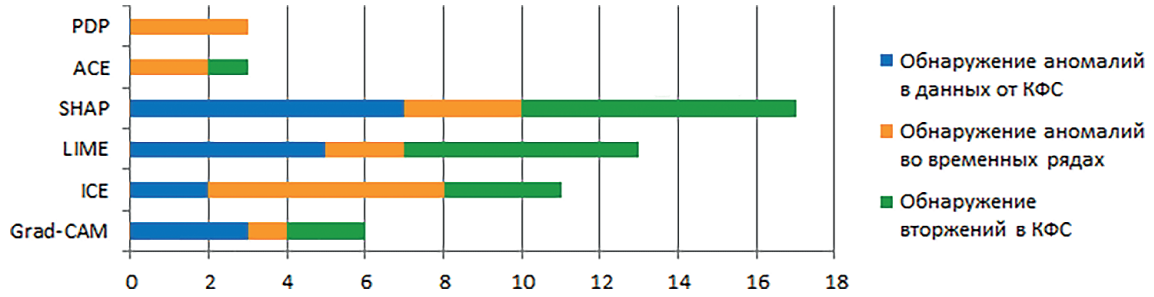


Рис. 1. Частота использования методов объяснения в различных задачах

для объяснения выхода моделей обнаружения аномалий. Следует отметить, что приведенные выше методы относятся к так называемым локальным методам объяснения, то есть позволяют объяснить отдельные предсказания для каждого экземпляра данных, но не поведение модели в целом, как это делают глобальные методы объяснения. В общем случае результатом локального метода объяснения является подмножество признаков данных, значения которых наиболее сильно повлияли на предсказание, сделанное используемой моделью машинного обучения.

Для генерации объяснений к выявленным аномалиям во временных рядах применяются методы, которые, в основном, были разработаны для текстов и изображений. В частности, в [9, 17–19] применяются методы LIME и SHAP. Амели и др. [20] предложили использовать методы на основе карт значимости. Для того, чтобы понять какие методы генерации объяснений наиболее часто используются для формирования объяснений аномалий во временных рядах, было проанализировано более 300 научных статей, извлеченных из электронной базы данных Elsevier Science Direct, опубликованных в интервале с 2021 по 2024 год по модели открытого доступа. Для поиска статей использовались следующие ключевые слова: «anomaly detection in CPS» (обнаружение аномалий в КФС), «anomaly detection in time series» (обнаружение аномалий во временных рядах), «intrusion detection» (обнаружение вторжений). На рис. 1 показана частота использования различных методов генерации объяснений в выбранных статьях (по оси абсцисс показано количество статей): GRAD-CAM, ICE (Individual Conditional expectation, индивидуальное условное ожидание), LIME, SHAP, Shapley values (значения Шепли), ALE (Accumulated Local Effects, накопленные локальные эффекты), PDP (Partial Dependence Plot, график частичной зависимости).

Подход к объяснимому обнаружению аномалий в потоке данных от технологического процесса

Предлагаемый подход к объяснимому обнаружению аномалий в потоке данных от технологического процесса состоит из следующих шагов (рис. 2):

1) предобработка входного потока данных, включающая преобразование входного вектора в матрицу;
2) выявление аномалий на основе сверточной нейронной сети;
3) генерация объяснений к прогнозам модели в виде вектора датчиков и актуаторов технологического процесса.

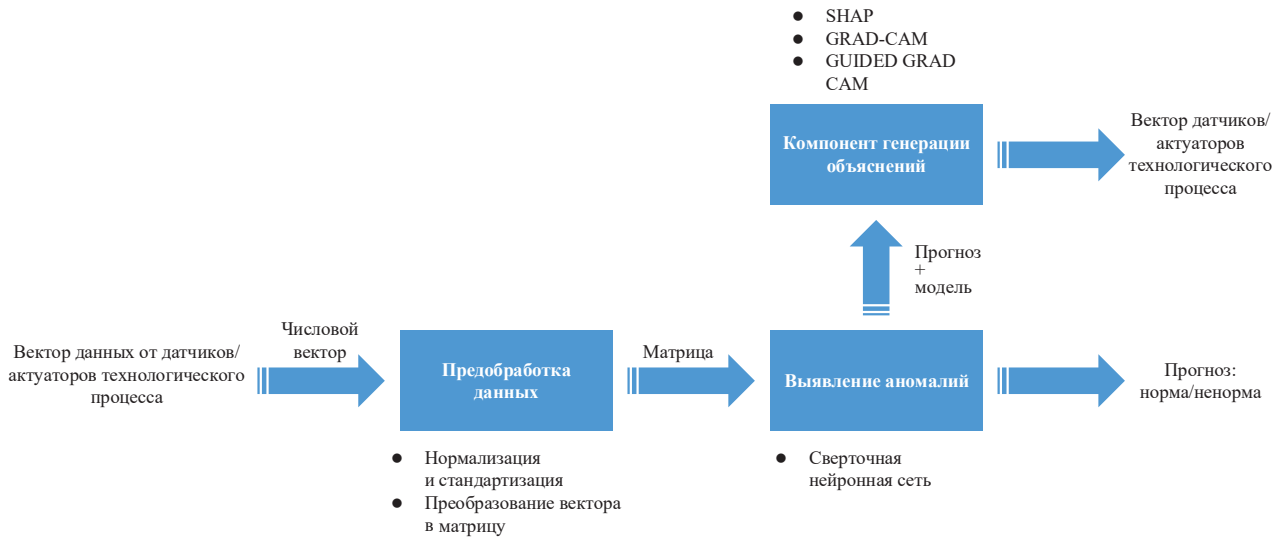


Рис. 2. Основные этапы объяснимого обнаружения аномалий в потоке данных от технологического процесса

Особенностью предлагаемого подхода является преобразование вектора данных в матрицу и использование сверточных нейронных сетей для выявления аномалий, что позволяет использовать методы генерации объяснений, специально разработанные для нейронных сетей данной архитектуры.

Преобразование входного вектора в матрицу выполняется в несколько шагов. Сначала значения входного вектора нормализуются, а затем компонуются в матрицу $n \times n$. Определение размеров матрицы осуществляется по формуле, предложенной в [3]:

$$n = \text{ceil}\left((K + N)/2\right),$$

где K – число количественных атрибутов, а N – число значений, которые могут принимать все категориальные атрибуты, ceil – функция округления в большую сторону до ближайшего целого числа.

Компоновка атрибутов в матрицу может быть осуществлено двумя способами. Первый способ основан только на последовательности атрибутов во входном векторе и не учитывает их подобия друг с другом. Строки матрицы заполняются последовательно, а неиспользованные элементы заполняются нулями. Такой способ часто называется прямой компоновкой данных [3]. В основе второго способа лежит идея, что атрибуты, которые подобны друг другу, должны располагаться в матрице ближе друг к другу, и в этом случае генерируемая матрица отражает пространственные закономерности в данных [21–23].

Подобие между атрибутами чаще всего вычисляется на основе оценки попарного сходства атрибутов, которая может быть представлена косинусным расстоянием, евклидовым расстоянием [24], коэффициентом корреляции Пирсона [25, 26] и т.д. Матрица попарного расстояния служит основой для упорядочивания атрибутов в исходной матрице [23].

В [21, 22] предложен другой подход к упорядочиванию атрибутов в матрице – подобие атрибутов устанавливается на основе построения их проекции в двумерном пространстве. Для нахождения проекции атрибутов сначала выполняется транспонирование обучающей выборки данных, в этом случае каждый атрибут описывается многомерным вектором, а далее могут быть применены как линейные, так и нелинейные алгоритмы снижения размерности. В данной работе анализируются два разных способа построения изображения: прямое преобразование и нелинейное преобразование DeepInsight на основе алгоритма t-SNE.

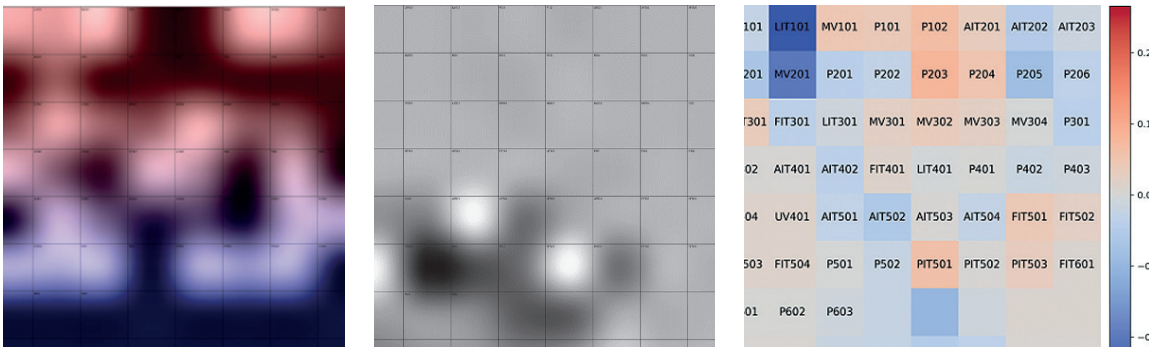
Формат входных данных в виде матрицы позволяет выбирать архитектуры нейронных сетей, в которых сверточные слои используются для извлечения анализируемых признаков. В настоящей работе в качестве аналитической модели предложено использовать простую двуслойную сверточную сеть.

Для генерации объяснений прогнозов сверточной сети предложено исследовать несколько подходов: метод SHAP и методы, разработанные специально для сверточных нейронных сетей Grad-CAM, Guided Grad-CAM.

Метод SHAP не зависит от архитектуры анализируемой модели и применим как для табличных данных, так и для изображений. В его основе лежит теория кооперативных игр, что позволяет оценить вклад каждого признака в конечное решение модели. В контексте решаемой задачи SHAP может быть использован для выявления пикселей, которые наиболее значимы для принятия решения.

Методы Grad-CAM и Guided Grad-CAM основаны на вычислении градиентов выхода модели относительно карт признаков последнего сверточного слоя. Метод Grad-CAM позволяет получить тепловую карту, подсвечивающую важные области входного изображения. Метод Guided Grad-CAM строит более детализированные и точные объяснения, благодаря комбинированию методов Grad-CAM и метода управляемого обратного распространения ошибки (guided backpropagation).

На рис. 3 представлены примеры объяснений, генерируемых разными способами.



а) б) в)

Рис. 3. Примеры генерируемых тепловых карт разными методами: а) методом Grad-CAM, б) методом Guided Grad-CAM, в) методом SHAP

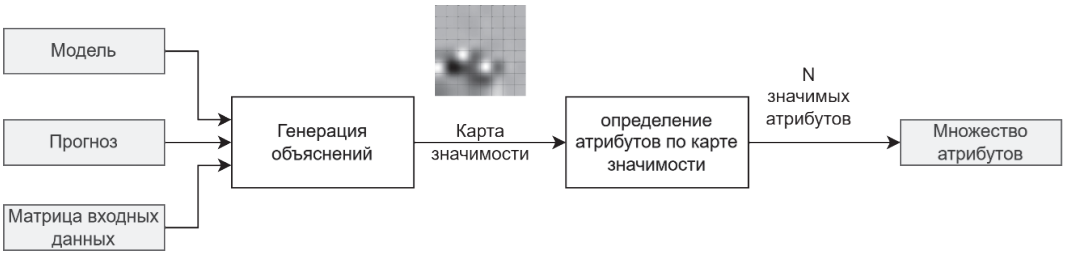


Рис. 4. Общая схема генерации объяснений различными методами

Для метода Grad-CAM значимость признака определяется по цвету: красный – признаки с высокой значимостью, синий – с низкой. В Guided Grad-CAM выполняется выделение областей значимых областей – чем они темнее, тем больший вклад они оказывают на выполненное аналитической моделью решение. В методе SHAP «красные» признаки указывают на то, что признак увеличивает вероятность предсказания, в то время как признаки, отмеченные синим цветом, наоборот, уменьшают эту вероятность. Таким образом, для получения списка потенциально аномальных данных необходимо выполнить обратную операцию нахождения атрибута по матрице. Общая схема генерации объяснений для каждого из метода представлена на рис. 4.

Для выполнения экспериментальной оценки была разработана библиотека DatasetToImageTransformer[6] на языке Python, которая выполняет преобразование числовых данных в матрицы (изображений) разными способами. Для каждого способа построения матрицы формируется словарь позиций атрибутов, в котором каждому элементу матрицы ставится в соответствие название атрибутов. Такое решение позволяет использовать данную библиотеку для анализа потока данных, а также выполнять обратное преобразование – по координатам элемента матрицы получить название атрибута. Для размеченных данных библиотека создает обучающие наборы данных, сгруппированные по каталогам, которые могут быть использованы для обучения различных аналитических моделей.

Экспериментальная оценка предложенного подхода

Целью эксперимента являлось определение эффективности предложенного подхода к выявлению аномалий, который состоит из одинаково значимых задач – обнаружения аномалий и их объяснений. Сценарий эксперимента был разработан таким образом, чтобы оценить эффективность каждого компонента подхода. В первой части эксперимента выполнялась оценка эффективности обнаружения аномальных данных от технологического процесса, на второй части эксперимента – оценка эффективности методов генерации объяснений.

6 https://github.com/Kotolow/FTIConverter.git

В обоих случаях использовался набор данных Secure Water Treatment (SWaT) версии 2015 [27], созданный с помощью программно-аппаратного стенда, представляющего собой уменьшенную копию водоочистных сооружений. Данный набор отражает 11 дней функционирования системы, 7 дней из которых соответствуют норме, а 4 дня содержат 36 атак разной длительности. В таблице 1 представлены характеристики набора SWaT версии 2015 года для 4-х аномальных дней, включая число датчиков, значения которых были изменены.

Таблица 1.
Характеристика атак в наборе данных SWaT

Тип записи в наборе	Число датчиков, измененных в результате атак	Число записей
Норма	0	399157
Аномалия	1	43213
	2	7789
	3	2452

Атаки отличаются числом атакуемых датчиков. В частности, было выполнено 6 атак, целью которых была подмена значений двух и трех датчиков, а также 7 атак на датчики, принадлежащие разным технологическим подпроцессам. Кроме того, отличительной особенностью данного набора является наличие текстовых объяснений, какие вредоносные воздействия проводились со стендом, и какие датчики и актуаторы были изменены, что делает его пригодным для оценки эффективности методов генерации объяснений.

На первом этапе в качестве метрик эффективности обнаружения аномалий были использованы показатели точности (precision), полноты (recall) и F1-мера, которые вычисляются на основе матрицы ошибок. Результаты экспериментов для разных способов генерации матрицы представлены в таблице 2. Из нее следует, что способ построения входной матрицы не влияет на точность обнаружения аномалий в потоке данных. Применение достаточно простой сверточной нейронной сети дает высокую точность решения задачи.

Таблица 2.

Оценка эффективности обнаружения аномалий

Способ генерации матрицы	Прямое преобразование			Нелинейное преобразование DeepInsight на основе t-SNE			Число записей
Класс	Точность (precision)	Полнота (recall)	F1-мера	Точность (precision)	Полнота (recall)	F1-мера	
Аномалия (атака)	0,99	0,94	0,96	0,98	0,93	0,96	54621
Норма	0,99	1,00	0,99	0,99	1,00	0,99	395298
Макро среднее	0,99	0,97	0,98	0,99	0,96	0,98	449919
Микро среднее	0,99	0,99	0,99	0,99	0,99	0,99	449919

На втором этапе эксперимента была выполнена оценка эффективности компонента генерации объяснений, которая определялась как точность объяснений. В данной работе точность объяснений предлагается оценивать на основе сравнения множества датчиков/актуаторов, которые были реально изменены в ходе деструктивных воздействий на систему, со множеством датчиков/актуаторов, которое было получено в результате применения методов генерации объяснений для каждого прогноза модели. В этих целях предложена метрика AHR (Any Hit Rate)[7], которая вычисляется следующим образом.

Пусть E_i есть множество датчиков/актуаторов, которые демонстрируют аномальное поведение в i-й момент времени, т.е. определены для i-ой точки данных, а E_i^* – множество датчиков/актуаторов, которые были получены в результате применения метода генерации к i-му прогнозу. Тогда

$$AHR = \frac{\sum_{i=1}^{N} any\_hit(i, max\_overlap)}{N},$$

$$\text{где } any\_hit(i) = \begin{cases} 1, E_i \cap E_i^* \neq \emptyset; \\ 0, E_i \cap E_i^* = \emptyset \end{cases},$$

где функция $any\_hit(i, max\_overlap)$ возвращает 1, если число совпадающих датчиков между объяснением и реальными данными больше или равно порогового значения $max\_overlap$, и 0 – в противном случае. Метрика AHR обладает высокой практической значимостью для промышленных систем, так как даже частичная локализация аномального поведения позволит специалистам выявить причину аномалии и своевременно принять необходимые контрмеры.

Очевидно, что для практического применения данной метрики необходимо выполнить преобразование исходного набора данных, дополнив его данными от аномальных сенсоров. В таблице 3 представлен пример измененного набора данных.

Следует также отметить, что из анализа были исключены записи, для которых не были указаны аномальные датчики; примером такой записи служит последняя строка в таблице 3, в которой «[]» обозначают пустой массив аномальных датчиков.

Также исходное множество было разбито на три подмножества:

- TP – строки из подмножества, для которых реально определено атакующее воздействие, которое было верно определено бинарным классификатором;
- $TP + FN$ – все строки из подмножества, для которых реально определено атакующее воздействие;
- $TP + FN + FP$ – все строки из подмножества, для которых реально определено атакующее воздействие, и строки, для которых детектор аномалий ошибочно предсказал состояние «аномалия».

Таблица 3.

Фрагмент измененного набора данных SWaT

Временная метка	FIT101	LIT101	MV101	...	Датчики
2015-12-28 10:28:14	2.494	817.674	2	...	[MV101]
2015-12-28 10:28:15	2.536	817.974	2	...	[MV101, P205]
...	...	...	...	...	...
2016-01-01 10:28:14	2.420	573.522	2	...	[]

7 https://www.researchgate.net/publication/360076778_Unsupervised_Multi-Sensor_Anomaly_Localization_with_Explainable_AI

Таблица 4.

Точность сформированных объяснений различными методами генерации объяснений

Метод трансформации вектора данных	Методы генерации объяснений	*AHR* для множества *TP*	*AHR* для множества *TP + FN*	*AHR* для множества *TP + FN + FP*
Нелинейное преобразование DeepInsight на основе t-SNE	Grad-CAM	0,2120	0,0154	0,0049
	Guided Grad-CAM	0,0372	0,0904	0,0009
	SHAP	0,0263	0,0034	0,0002
Прямое преобразование	Grad-CAM	0,0678	0,0466	0,0007
	Guided Grad-CAM	0,1296	0,0718	0,0014
	SHAP	0,0667	0,5973	0,0005

Полученные результаты представлены в таблице 4.

Очевидно, что все методы дают крайне низкую точность, неприемлемую для случаев практического использования. Так, например, метод Grad-CAM достигает максимальной точности для нелинейного метода построения матрицы DeepInsight на множестве верно выявленных аномалий и составляет 0,21. Метод Guided Grad-CAM достигает максимальной точности на множестве верно выявленных аномалий для прямого преобразования. Метод SHAP демонстрирует максимальную точность на множестве $TP + FN$, т.е. на множестве, на котором определены реальные атакующие воздействия. Его точность вначале составляет 0,5973, но резко падает до 0,0005 на множестве векторов, которые включают вектора, которые классификатор относит к аномальным. Это делает неприемлемым использование и данного метода на практике. Возможной причиной такой низкой точности является природа как самих анализируемых данных – временные ряды, так и самих аномалий, которые имеют длительность и могут выражаться различной степенью изменения атрибутов. Предложенные преобразования над входными данными и сама модель учитывают только пространственные связи между данными, при этом временные зависимости между ними не учитываются. Между тем, упомянутые выше методы основаны на предположении, что атрибуты между собой независимы, а данные не зависят друг от друга, что неверно для набора SWaT. Таким образом, выявлена острая необходимость в разработке методов генерации объяснений для моделей машинного обучения, предназначенных для временных рядов.

Заключение

В работе представлен подход к объяснимому выявлению аномалий в потоке данных от технологических процессов. Отличительной особенностью предложенного подхода является использование преобразования входного вектора данных в матрицу, что позволяет применять сверточные нейронные слои для извлечения анализируемых признаков. Выявление аномалий осуществляется при помощи двухслойной сверточной сети, которая показала высокую точность обнаружения аномалий для тестируемого набора данных SWAT, описывающего функционирование системы водоочистных сооружений.

Для реализации компонента формирования объяснений было использовано несколько методов генерации объяснений: методы Grad-CAM, Guided Grad-CAM и SHAP. Было показано, что на текущий момент точность формируемых объяснений низкая, что делает невозможным применение данных методов на практике. Возможной причиной низкой точности являются сами исследуемые данные – многомерные временные ряды.

Дальнейшие направления исследований по этой задаче связаны с разработкой новых моделей выявления аномалий, которые учитывают не только пространственные связи между атрибутами, но и временные. Возможным решением служит использование графовых нейронных сетей. Кроме того, планируется исследование и разработка методов генерации объяснений, которые учитывают особенности многомерных временных рядов, а именно наличие связей между атрибутами как во времени, так и между собой.

Благодарность. Исследование выполнено при поддержке гранта Российского научного фонда №23-11-20024, https://rscf.ru/project/23-11-20024/, и Санкт-Петербургского научного фонда в СПб ФИЦ РАН.

Рецензент: Лаута Олег Сергеевич, доктор технических наук, профессор кафедры комплексного обеспечения информационной безопасности Государственного университета морского и речного флота имени адмирала С. О. Макарова, Санкт-Петербург, Россия. E-mail: laos-82@yandex.ru

Литература

1. 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.
2. Котенко И. В., Федорченко Е. В., Новикова Е. С., Саенко И. Б., Данилов А. С. Методология сбора данных для анализа безопасности промышленных киберфизических систем // Вопросы кибербезопасности. 2023. № 5(57). С. 69–79. https://doi.org/10.21681/2311-3456-2023-5-69-79.
3. Novikova E. S., Fedorchenko E. V., Bukhtiyarov M. A., Saenko I. B. Anomaly detection in wastewater treatment process for cyber resilience risks evaluation // Journal of Mining Institute. 2024. Vol. 267. P. 488–500.
4. Dong H., Kotenko I. Cybersecurity in the AI era: analyzing the impact of machine learning on intrusion detection // Knowledge and Information Systems, 2025, 67(5), P. 3915–3966, 102748. DOI: 10.1007/s10115-025-02366-w.
5. Kotenko I. V., Levshun D. A. Machine Learning Methods of Intelligent System Event Analysis for Multistep Cyberattack Detection // Scientific and Technical Information Processing, 2024, Vol. 51, No. 5, P.372–381. Allerton Press, Inc., 2024. Springer Nature. ISSN 0147-6882. DOI: 10.3103/S0147688224700254.
6. Dong H., Kotenko I., Levshun D. Next-Generation IIoT Security: Comprehensive Comparative Analysis of CNN-based Approaches // Knowledge Based Systems, Vol.316, 12 May 2025, 113337. https://doi.org/10.1016/j.knosys.2025.113337.
7. Doynikova E., Novikova E., Murenin I., Kolomeec M., Gaifulina D., Tushkanova O., Levshun D., Meleshko A., Kotenko I. Security Measuring System for IoT Devices // Lecture Notes in Computer Science. 2022. Vol. 13106. P. 256–275.
8. Ning X., Jiang J. Design, Analysis and Implementation of a Security Assessment/Enhancement Platform for Cyber-Physical Systems // IEEE Transactions on Industrial Informatics. 2022. Vol. 18. No. 2. P. 1154–1164.
9. Wang C., Wang B., Liu H., Qu H. Anomaly detection for industrial control system based on autoencoder neural network // Wirel. Commun. Mob. Comput. 2020. P. 8897926–1889792610.
10. Rodríguez M., Tobón D., Múnera D. A framework for anomaly classification in Industrial Internet of Things systems // Internet of Things. 2025. Vol. 29. Article 101446. https://doi.org/10.1016/j.iot.2024.101446.
11. Su Y., Zhao Y., Niu C., Liu R., Sun W., Pei D. Robust anomaly detection for multivariate time series through stochastic recurrent neural network // Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '19). ACM, New York, NY, USA, 2019, pp. 2828–2837. https://doi.org/10.1145/3292500.3330672.
12. Nizam H., Zafar S., Lv Z., Wang F., Hu X. Real-Time Deep Anomaly Detection Framework for Multivariate Time-Series Data in Industrial IoT // IEEE Sensors Journal. 2022. Vol. 22. No. 23. P. 22836–22849, doi: 10.1109/JSEN.2022.3211874.
13. Liu Y. et al. Deep Anomaly Detection for Time-Series Data in Industrial IoT: A Communication-Efficient On-Device Federated Learning Approach // IEEE Internet of Things Journal. 2021. Vol. 8. No. 8. P. 6348–6358. doi: 10.1109/JIOT.2020.3011726.
14. Zhao P., Ding Z., Li Y., Zhang X., Zhao Y., Wang H, Yang Y. SGAD-GAN: Simultaneous Generation and Anomaly Detection for time-series sensor data with Generative Adversarial Networks // Mechanical Systems and Signal Processing. 2024. Vol. 210. Article 111141. https://doi.org/10.1016/j.ymssp.2024.111141.
15. Lundberg S. M., Lee S.-I. A unified approach to interpreting model predictions // Advances in neural information processing systems (NIPS'17), 2017, pp. 4768–4777.
16. Ribeiro M. T., Singh S., Guestrin C. Why Should I Trust You?: Explaining the Predictions of Any Classifier // Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'16). ACM, NY, USA, 2016, pp. 1135–1144.
17. Neshenko N., Bou-Harb E., Furht B. A behavioral-based forensic investigation approach for analyzing attacks on water plants using GANs // Forensic Science International: Digital Investigation. 2021. Vol. 37. Article 301198.
18. Antwarg L., Miller R. M., Shapira B., Rokach L. Explaining anomalies detected by autoencoders using SHAP. arXiv preprint arXiv:1903.02407. 2019.
19. Oliveira D., Vismari L. F., Nascimento A. M., de Almeida J. R., Cugnasca P. S., Camargo J. B., Almeida L., Gripp R., Neves M. A new interpretable unsupervised anomaly detection method based on residual explanation // IEEE Access. 2021. Vol. 10, pp. 1401–1409.
20. Ameli M., Becker P. A., Lankers K., van Ackeren M., Bähring H., Maaß W. Explainable unsupervised multi-sensor industrial anomaly detection and categorization // 21st IEEE International Conference on Machine Learning and Applications (ICMLA), 2022, pp. 1468–1475.
21. Sharma A., Vans E., Shigemizu D., Boroevich K. A., Tsunoda T. DeepInsight: A methodology to transform a non-image data to an image for convolution neural network architecture. Sci. Rep. 2019. Vol. 9. Article 11399. https://doi.org/10.1038/s41598-019-47765-6.
22. Bazgir O., Zhang R., Dhruba S. R., Rahman R., Ghosh S., Pal R. Representation of features as images with neighborhood dependencies for compatibility with convolutional neural networks. Nat. Commun. 2020. Vol. 11. Article 4391. https://doi.org/10.1038/s41467-020-18197-y.
23. Zhu Y., Brettin T., Xia F., Partin A., Shukla M., Yoo H., Evrard Y. A., Doroshow J. H., Stevens R. L. Converting tabular data into images for deep learning with convolutional neural networks. Sci. Rep. 2021. Vol. 11. Article 11325. https://doi.org/10.1038/s41598-021-90923-y.
24. Zhou Q., Chen J., Liu H., He S., Meng W. Detecting Multivariate Time Series Anomalies with Zero Known Label. 2022. arXiv.org/abs/2208.02108.
25. Xie Y., Zhang H., Babar M. A. Multivariate Time Series Anomaly Detection by Capturing Coarse-Grained Intra- and Inter-Variate Dependencies. 2025. arXiv.org/abs/2501.16364.
26. Kamarthi H., Kong L., Rodriguez A., Zhang C., Prakash B. A. Learning Graph Structures and Uncertainty for Accurate and Calibrated Time-series Forecasting. 2024. arXiv.org/abs/2407.02641.
27. Goh J., Adepu S., Junejo K., Mathur A. A Dataset to Support Research in the Design of Secure Water Treatment Systems // Critical Information Infrastructures Security. CRITIS 2016. Lecture Notes in Computer Science. Vol. 10242. Springer, Cham. https://doi.org/10.1007/978-3-319-71368-7_8.

AN APPROACH TO EXPLAINABLE ANOMALY DETECTION IN DATA STREAMS FROM TECHNOLOGICAL SYSTEMS

Novikova E. S.[8], Bukhtiarov M. A.[9], Kotenko I. V.[10], Saenko I. B.[11], Fedorchenko E. V.[12]

Keywords: *cyber attack and anomaly detection, industrial cyberphysical systems anomaly generation, evaluation of explanation accuracy.*

The purpose of the study: *development of an approach to identify anomalies in process data based on explainable machine learning in order to further select countermeasures taking into account possible sources of anomalies.*

Research methods: *statistical analysis, machine learning methods, methods of generating explanations for machine learning model predictions.*

Results obtained: *an approach to explainable anomaly detection in the flow of data from technological processes is proposed, its main stages are presented, which is based on the transformation of the input data vector into a matrix, and the detection of anomalies using a convolutional neural network; the method of transformation of the data vector into a matrix is developed and the influence of the data transformation algorithm on the efficiency of solving the problem of anomaly detection is evaluated; the method of testing the accuracy of the generated explanations is developed and the experimental evaluation is carried out.*

Scientific novelty: *the proposed approach to the identification of anomalies in process data differs from the existing ones by using the technique of transforming the input data vector into a matrix, which allows us to apply a convolutional neural network as an analytical model of anomaly detection and methods of generating explanations developed specifically for neural networks of this architecture.*

Contributions: *Evgenia Novikova – development of a method for converting the input data flow; Marat Bukhtiyarov – experimental study of the proposed approach; Igor Kotenko – development of a general approach to explainable detection of anomalies of the concept of dynamic assessment of the security of information systems in conditions of uncertainty of the initial data; Igor Kotenko, Igor Saenko and Elena Fedorchenko – analysis of the state of arts in identifying anomalies in technological processes and forming explanations for forecasts of machine learning models.*

References

1. Levshun D. A., Levshun D. S., Kotenko I. V. Detecting and explaining anomalies in industrial Internet of things systems using an autoencoder // Ontology of designing. 2025. Vol.15, No.1(55). P.96-113. DOI:10.18287/2223-9537-2025-15-1-96-113.
2. Kotenko I. V., Fedorchenko E. V., Novikova E. S., Saenko I. B., Danilov A. S. Methodology of data collection for security analysis of industrial cyber-physical systems // Cybersecurity Issues. 2023. No. 5 (57). P. 69-79. https://doi.org/10.21681/2311-3456-2023-5-69-79.
3. Novikova E. S., Fedorchenko E. V., Bukhtiyarov M. A., Saenko I. B. Anomaly detection in wastewater treatment process for cyber resilience risks evaluation // Journal of Mining Institute. 2024. Vol. 267. P. 488–500.
4. Dong H., Kotenko I. Cybersecurity in the AI era: analyzing the impact of machine learning on intrusion detection // Knowledge and Information Systems, 2025, 67(5), P. 3915–3966, 102748. DOI: 10.1007/s10115-025-02366-w.
5. Kotenko I. V., Levshun D. A. Machine Learning Methods of Intelligent System Event Analysis for Multistep Cyberattack Detection // Scientific and Technical Information Processing, 2024, Vol. 51, No. 5, P.372–381. Allerton Press Inc., 2024. Springer Nature. ISSN 0147-6882. DOI: 10.3103/S0147688224700254
6. Dong H., Kotenko I., Levshun D. Next-Generation IIoT Security: Comprehensive Comparative Analysis of CNN-based Approaches // Knowledge Based Systems, Vol.316, 12 May 2025, 113337. https://doi.org/10.1016/j.knosys.2025.113337.
7. Doynikova E., Novikova E., Murenin I., Kolomeec M., Gaifulina D., Tushkanova O., Levshun D., Meleshko A., Kotenko I. Security Measuring System for IoT Devices // Lecture Notes in Computer Science. 2022. Vol. 13106. P. 256–275.
8. Ning X., Jiang J. Design, Analysis and Implementation of a Security Assessment/Enhancement Platform for Cyber-Physical Systems // IEEE Transactions on Industrial Informatics. 2022. Vol. 18. No. 2. P. 1154–1164.
9. Wang C., Wang B., Liu H., Qu H. Anomaly detection for industrial control system based on autoencoder neural network // Wirel. Commun. Mob. Comput. 2020. P. 8897926–1889792610.
10. Rodríguez M., Tobón D., Múnera D. A framework for anomaly classification in Industrial Internet of Things systems // Internet of Things. 2025. Vol. 29. Article 101446. https://doi.org/10.1016/j.iot.2024.101446.

8 Evgenia S. Novikova, Ph.D. of Technical Sciences, Senior researcher of Laboratory of Computer Security Problems at St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS), St. Petersburg, Russia. E-mail: novikova@comsec.spb.ru

9 Marat A. Bukhtiarov, Software Developer of Laboratory of Computer Security Problems at St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS), St. Petersburg, Russia. E-mail: buhtiarov.marat@gmail.com

10 Igor V. Kotenko, Honored Worker of Science of the Russian Federation, Dr.Sc. of Technical Sciences, Professor, Chief Scientist and Head of Laboratory of Computer Security Problems at St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS), St. Petersburg, Russia. E-mail: ivkote@comsec.spb.ru

11 Igor B. Saenko, Dr.Sc. of Technical Sciences, Professor, Leading researcher of Laboratory of Computer Security Problems at St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS), St. Petersburg, Russia. E-mail: ibsaen@comsec.spb.ru

12 Elena V. Fedorchenko, Ph.D. of Technical Sciences, Senior researcher of Laboratory of Computer Security Problems at St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS), St. Petersburg, Russia. E-mail: doynikova@comsec.spb.ru

11. Su Y., Zhao Y., Niu C., Liu R., Sun W., Pei D. Robust anomaly detection for multivariate time series through stochastic recurrent neural network // Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '19). ACM, New York, NY, USA, 2019, pp. 2828–2837. https://doi.org/10.1145/3292500.3330672.
12. Nizam H., Zafar S., Lv Z., Wang F., Hu X. Real-Time Deep Anomaly Detection Framework for Multivariate Time-Series Data in Industrial IoT // IEEE Sensors Journal. 2022. Vol. 22. No. 23. P. 22836–22849, doi: 10.1109/JSEN.2022.3211874.
13. Liu Y. et al. Deep Anomaly Detection for Time-Series Data in Industrial IoT: A Communication-Efficient On-Device Federated Learning Approach // IEEE Internet of Things Journal. 2021. Vol. 8. No. 8. P. 6348–6358. doi: 10.1109/JIOT.2020.3011726.
14. Zhao P., Ding Z., Li Y., Zhang X., Zhao Y., Wang H, Yang Y. SGAD-GAN: Simultaneous Generation and Anomaly Detection for time-series sensor data with Generative Adversarial Networks // Mechanical Systems and Signal Processing. 2024. Vol. 210. Article 111141. https://doi.org/10.1016/j.ymssp.2024.111141.
15. Lundberg S. M., Lee S. -I. A unified approach to interpreting model predictions // Advances in neural information processing systems (NIPS'17), 2017, pp. 4768–4777.
16. Ribeiro M. T., Singh S., Guestrin C. Why Should I Trust You?: Explaining the Predictions of Any Classifier // Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'16). ACM, NY, USA, 2016, pp. 1135–1144.
17. Neshenko N., Bou-Harb E., Furht B. A behavioral-based forensic investigation approach for analyzing attacks on water plants using GANs // Forensic Science International: Digital Investigation. 2021. Vol. 37. Article 301198.
18. Antwarg L., Miller R. M., Shapira B., Rokach L. Explaining anomalies detected by autoencoders using SHAP. arXiv preprint arXiv:1903.02407. 2019.
19. Oliveira D., Vismari L. F., Nascimento A. M., de Almeida J. R., Cugnasca P. S., Camargo J. B., Almeida L., Gripp R., Neves M. A new interpretable unsupervised anomaly detection method based on residual explanation // IEEE Access. 2021. Vol. 10, pp. 1401–1409.
20. Ameli M., Becker P. A., Lankers K., van Ackeren M., Bähring H., Maaß W. Explainable unsupervised multi-sensor industrial anomaly detection and categorization // 21st IEEE International Conference on Machine Learning and Applications (ICMLA), 2022, pp. 1468–1475.
21. Sharma A., Vans E., Shigemizu D., Boroevich K. A., Tsunoda T. DeepInsight: A methodology to transform a non-image data to an image for convolution neural network architecture. Sci. Rep. 2019. Vol. 9. Article 11399. https://doi.org/10.1038/s41598-019-47765-6.
22. Bazgir O., Zhang R., Dhruba S. R., Rahman R., Ghosh S., Pal R. Representation of features as images with neighborhood dependencies for compatibility with convolutional neural networks. Nat. Commun. 2020. Vol. 11. Article 4391. https://doi.org/10.1038/s41467-020-18197-y.
23. Zhu Y., Brettin T., Xia F., Partin A., Shukla M., Yoo H., Evrard Y. A., Doroshow J. H., Stevens R. L. Converting tabular data into images for deep learning with convolutional neural networks. Sci. Rep. 2021. Vol. 11. Article 11325. https://doi.org/10.1038/s41598-021-90923-y.
24. Zhou Q., Chen J., Liu H., He S., Meng W. Detecting Multivariate Time Series Anomalies with Zero Known Label. 2022. arXiv.org/abs/2208.02108.
25. Xie Y., Zhang H., Babar M. A. Multivariate Time Series Anomaly Detection by Capturing Coarse-Grained Intra- and Inter-Variate Dependencies. 2025. arXiv.org/abs/2501.16364.
26. Kamarthi H., Kong L., Rodriguez A., Zhang C., Prakash B. A. Learning Graph Structures and Uncertainty for Accurate and Calibrated Time-series Forecasting. 2024. arXiv.org/abs/2407.02641.
27. Goh J., Adepu S., Junejo K., Mathur A. A Dataset to Support Research in the Design of Secure Water Treatment Systems // Critical Information Infrastructures Security. CRITIS 2016. Lecture Notes in Computer Science. Vol. 10242. Springer, Cham. https://doi.org/10.1007/978-3-319-71368-7_8.