УДК 004.8 **DOI: 10.18413/2518-1092-2026-11-1-0-9**

Котов Д.В.

КЛАССИФИКАЦИЯ АНОМАЛИЙ В ДАННЫХ И ВЫБОР МЕТОДОВ ИХ ОБНАРУЖЕНИЯ

Военная академия Генерального штаба Вооружённых си-л Российской Федерации
пр-т Вернадского, 100, г. Москва, 119571, Россия

e-mail: kotovdv2101@outlook.com

Аннотация
Выявление аномалий стало базовой задачей анализа данных в условиях роста объёмов наблюдений, усложнения информационных систем и перехода к мониторингу в реальном времени. Аномальные события часто редки, но именно они связаны с наиболее значимыми рисками и эффектами: отказами оборудования, мошенничеством или кибератаками. Отклонение может быть абсолютным (выброс), условным (аномально только в конкретном контексте) или проявляться на уровне группы наблюдений. Ошибки в типизации аномалий приводят к неверному выбору алгоритма, некорректной настройке порогов и, как следствие, к ложным срабатываниям или пропускам редких событий. В работе систематизированы три базовых типа аномалий – глобальные (точечные), контекстные и коллективные. Показано, как аномалии соотносятся с тремя ключевыми подходами к их детектированию: статистической, плотностно-дистанционной и модельной. Для каждого типа обсуждаются типичные алгоритмы (Z-score, IQR, Mahalanobis, kNN/LOF, Isolation Forest, кластеризация, модели временных рядов и последовательностей, автоэнкодеры, вариационные модели), а также требования к данным и вычислительным ресурсам. Предложена практическая схема выбора методов в зависимости от типа аномалии и модальности данных (табличные данные, временные ряды, потоки), приведены рекомендации по настройке порогов и по оценке качества на несбалансированных выборках (PR-AUC, MCC, событийные метрики). Классификация аномалий является необходимым этапом проектирования системы мониторинга; наилучшие результаты в прикладных задачах обычно достигаются каскадными и ансамблевыми решениями, сочетающими интерпретируемые базовые методы и гибкие модели машинного обучения.

Ключевые слова: аномалии данных; обнаружение аномалий; выбросы; контекст; временные ряды; автоэнкодер; Isolation Forest; LOF

Для цитирования: Котов Д.В. Классификация аномалий в данных и выбор методов их обнаружения // Научный результат. Информационные технологии. – Т.11, №1, 2026. – С. 106-116. DOI: 10.18413/2518-1092-2026-11-1-0-9

Kotov D.V.

DATA ANOMALY TAXONOMY AND METHOD SELECTION

Military Academy of the General Staff of the Armed Forces of the Russian Federation,
100 Vernadsky Ave., Moscow, 119571, Russian

e-mail: kotovdv2101@outlook.com

Abstract
Anomaly detection has become a core component of modern data analysis due to the growth of data volumes, the increasing complexity of information systems, and the demand for real-time monitoring. Although anomalies are usually rare, they often correspond to high-impact events such as equipment failures, fraud, cyberattacks, and critical medical conditions. There is no universal notion of an anomaly: deviations can be absolute, context-dependent, or emergent at the level of groups and sequences. Misclassifying the anomaly type leads to inappropriate modeling assumptions, poorly calibrated thresholds, and a trade-off skewed toward false alarms or missed events. This paper reviews three major anomaly types - global (point), contextual, and collective – and relates them to three detection paradigms: statistical tests, density/distance-based methods, and

model-based approaches. For each type, we discuss representative algorithms (Z-score, IQR, Mahalanobis distance, kNN/LOF, Isolation Forest, clustering, time-series and sequence models, LSTM/autoencoders, and variational models), together with their data requirements and practical limitations. We provide a method-selection scheme aligned with data modality (tabular data, time series, streaming data), recommend threshold calibration strategies, and outline evaluation protocols for highly imbalanced settings (PR-AUC, MCC, event-based metrics). Correct anomaly typing is a prerequisite for effective monitoring; in applied scenarios, the most robust solutions are typically cascade and ensemble pipelines that combine interpretable baselines with flexible machine learning models.
Keywords: data anomalies; anomaly detection; outliers; context; time series; autoencoder; Isolation Forest; LOF
For citation: Kotov D.V. Data Anomaly Taxonomy and Method Selection // Research result. Information technologies. – T.11, №1, 2026. – P. 106-116. DOI: 10.18413/2518-1092-2026-11-1-0-9

ВВЕДЕНИЕ

В практическом анализе данных понятие аномалии используется как минимум в двух значениях. С одной стороны, аномалия – это ошибка измерения или сбой в канале передачи, который желательно исправить до построения модели. С другой стороны, аномалия может быть искомым сигналом: мошенническая транзакция, нетипичное поведение пользователя или ранний признак отказа оборудования [1]. В обоих случаях обнаружение аномалий является самостоятельной задачей, которая не сводится к поиску явных выбросов или к применению одного универсального алгоритма [2].

Современные наборы данных усложняют задачу детектирования аномалий в связи с тем, что, данные многомерны и гетерогенны, т.е. числовые признаки представлены совместно с категориальными, текстовыми и векторными представлениями. Нормальное поведение источников данных часто имеет несколько режимов (несколько кластеров или состояний), поэтому предположение о едином распределении оказывается неверным [3] Во временных рядах нормальность зависит от контекста (сезонность, тренды, смена режимов), и одинаковые значения могут быть нормальными в одном периоде и аномальными в другом. В реальных системах наблюдается нестационарность: со временем меняются данные, правила и организационные мероприятия, что требует адаптивных критериев и контроля качества в эксплуатации [4]. Выбор алгоритма обнаружения аномалий должен начинаться не с перечня моделей, а с типизации ожидаемых аномалий. В литературе наиболее распространена трехчастная типология: глобальные (точечные), контекстные и коллективные аномалии [5]. Эти типы различаются масштабом проявления, зависимостью от контекста и тем, какая информация является достаточной для корректного решения. Следовательно, различается и набор подходящих методов поиска аномалий в данных.

Целью работы является систематизация видов аномалий данных и методов их обнаружения с акцентом на практический выбор алгоритма. В статье последовательно рассматриваются: формальные определения трех типов аномалий; связь между типом аномалии и методом детектирования; методы классической статистики, машинного обучения и нейросетевые архитектуры, применимые к каждому типу; практические рекомендации по настройке порогов и оценке качества в условиях сильного дисбаланса классов.

МАТЕРИАЛЫ И МЕТОДЫ ИССЛЕДОВАНИЯ

Пусть задан набор наблюдений $X = \{x_i\}_{i=1}^{n}$, где каждое наблюдение x_i принадлежит пространству признаков $\mathbb{R}^d$ (или более общему пространству после кодирования категорий, текста и др.). Задача обнаружения аномалий в наиболее общем виде сводится к построению функции оценки аномальности $s: X \to \mathbb{R}$, где большие значения $s(x)$ соответствуют большему отклонению объекта от нормы. Затем вводится правило принятия решения, обычно пороговое: x объявляется

аномалией, если $s(x) > T$. В потоковых задачах порог T может зависеть от времени и режима ($T = T(t)$), а сама функция $s(x)$ может пересчитываться в режиме реального времени [6].

С практической точки зрения важно различать три сценария обучения: обучение «с учителем», когда доступны размеченные примеры аномалий; полунаблюдаемое обучение, когда модель обучается на норме и ищет отклонения; обучение «без учителя», когда разметки нет, аномалии редки и выделяются как отклонения. В прикладных проектах чаще всего используют второй и третий подходы, поскольку разметка аномалий дорогая, неполная и быстро устаревает [2].

ТИПОЛОГИЯ АНОМАЛИЙ: ГЛОБАЛЬНЫЕ, КОНТЕКСТНЫЕ, КОЛЛЕКТИВНЫЕ

1. Глобальные (точечные) аномалии. Глобальная аномалия – это отдельное наблюдение, существенно отклоняющееся от большинства данных без учета внешнего контекста, представляющее собой маловероятное значение. Для одномерного признака типовым формализмом является стандартизированное отклонение:

$$z = \frac{x - \mu}{\sigma}; \quad (1)$$

где μ и σ – оценки центра и масштаба нормальных данных.

При $|z| > t$ (обычно $t = 2.5 - 3$) значение рассматривают как выброс [5]. Для многомерных данных аналогичную роль играет расстояние Махаланобиса, учитывающее ковариации признаков [3]. Важно, что для глобальных аномалий достаточно сравнения с общим распределением или общей геометрией данных.

2. Контекстные аномалии. Контекстная аномалия – это наблюдение, которое является аномальным только при заданном контексте C (время, место, режим, категория). Формально представляется возможным разделять признаки на контекстуальные x_c и поведенческие x_b. Аномальность определяется условно, через отклонение x_b от типичного поведения при заданном x_c [4].

3. Коллективные аномалии. Коллективная аномалия – это группа наблюдений $S = \{x_t\}_{t=t_0}^{t_0+k}$, которая в совокупности образует нетипичный паттерн, хотя отдельные точки могут быть нормальными. Наиболее часто коллективные аномалии возникают во временных рядах и потоках: резкое изменение уровня, серия повторяющихся событий, устойчивый сдвиг распределения, нетипичная корреляция признаков на интервале времени [5]. В инженерных системах коллективные аномалии могут соответствовать переходу в деградированный режим, когда изменение выражено не одной точкой, а траекторией, и важно не только обнаружить факт отклонения, но и определить его длительность и динамику [3].

4. Пересечения типов. Типы аномалий не являются взаимоисключающими. Запись может быть одновременно глобальной и контекстной (например, необычно большая сумма платежа в необычное время), а также быть частью коллективной аномалии (серия подобных платежей) [2]. Практическая система детектирования аномалий должна обеспечивать многоуровневую интерпретацию, различающую глобальные отклонения (аномалии, необычные сами по себе), контекстные отклонения (аномалии, обусловленные конкретным контекстом) и коллективные отклонения (аномалии, проявляющиеся в последовательностях или группах).

МЕТОДЫ ОБНАРУЖЕНИЯ АНОМАЛИЙ

Содержательно большинство методов можно свести к трем подходам: статистической, плотностно-дистанционной и модельной. Статистические методы предполагают, что нормальные данные подчиняются некоторому распределению, и рассматривают аномалии как события из хвостов распределения. Их сильные стороны – простота, низкая вычислительную стоимость, настройка пороговых значений. Однако, статистические методы не являются наиболее подходящими методами при работе с данными, которые представлены в высоких размерностях [4].

Статистическая парадигма особенно полезна как первый слой каскада, когда требуется быстро отсечь грубые ошибки измерения и явно невозможные значения.

Методы на основе плотности и расстояний исходят из того, что нормальные наблюдения образуют плотные области, а аномалии находятся в разреженных регионах. К таким методам относят kNN и LOF: они оценивают относительную локальную плотность [7, 8]. Их преимущество – гибкость, поскольку не требуется задавать параметрический вид распределения. В то же время недостатками выступают высокая вычислительная сложность и чувствительность к гиперпараметрам (число соседей, радиус окрестности, пороги плотности) [9, 10]. На практике использование этих алгоритмов обязательно предполагает предварительную стандартизацию признаков, а желательно и снижение размерности, иначе вычисляемые расстояния теряют информативность.

В основе модельных методов лежит построение модели нормального поведения, а аномальность оценивается через ошибку прогноза или восстановления данных. К этой группе относятся регрессионные и вероятностные модели, модели состояний, а также нейросетевые архитектуры (LSTM, автоэнкодеры, вариационные автокодировщики) [11]. Такие подходы способны улавливать сложные зависимости, включая нелинейные и временные, но при этом возможна проблема переобучения. Поэтому модельные методы редко применяются самостоятельно: они чаще выступают вторым или третьим слоем в каскадных схемах, следуя за базовой фильтрацией данных.

1. Методы обнаружения глобальных (точечных) аномалий.

1.1. Робастные статистические критерии.

Помимо классического Z-score применяются критерии на основе квартилей (IQR), а также робастные оценки центра и масштаба (медиана, MAD), которые устойчивее к наличию выбросов [3, 12]. Робастность критична, когда в данных уже есть аномалии и они смещают среднее и стандартное отклонение. С точки зрения внедрения преимущество статистических критериев заключаются в минимальных вычислениях, простоте применения для построения мониторинговых порогов.

1.2. Многомерные расстояния и локальная плотность.

Для коррелированных признаков расстояние Махаланобиса учитывает ковариации и корректно отражает геометрию расположения точек данных [3, 13]. Когда данные имеют несколько режимов или кластеров разной плотности, полезны локальные критерии: kNN-оценка и локальный фактор выбросов LOF [13]. LOF сравнивает плотность точки с плотностью ее окружения, поэтому способен выявлять локальные аномалии внутри неоднородных данных. Практически важно выбирать метрику и масштабирование так, чтобы метод отражал структуру данных, а не единицы измерения признаков.

1.3. Методы изоляции и одноклассовые модели.

Isolation Forest строит ансамбль случайных деревьев. На каждом шаге дерево случайным образом выбирает признак и точку разделения, «изолируя» отдельные наблюдения. Аномалии при этом отделяются быстрее, т.е. они достигают конечных узлов на меньшей глубине, чем точки из плотных областей. One-Class SVM строит границу, отделяющую нормальные данные, и часто используется как базовый полунаблюдаемый метод при наличии чистого обучающего набора, представляющего «норму» [14]. Оба подхода эффективны для табличных данных: они не опираются на предположение о нормальности распределения, работают быстрее методов на основе kNN и формируют единую оценку аномальности $s(x)$, пригодную для последующей калибровки.

1.4. Нейросетевые методы реконструкции и гибридные модели.

Автоэнкодер восстанавливает входной сигнал через промежуточное сжатое представление; степень отклонения от нормы оценивается по ошибке реконструкции (например, евклидово расстояние). Вариационные автоэнкодеры (VAE) вносят вероятностную трактовку: аномальность оценивается либо через вероятность восстановления наблюдения, либо через степень отклонения в пространстве скрытых переменных. Для сложных, многомодальных распределений применяют гибридные решения, такие как DAGMM (Deep Autoencoding Gaussian Mixture Model), где

автоэнкодер сочетается с кластеризацией, что повышает устойчивость к разнородности данных [11, 15]. При внедрении нейросетевых методов критически важно корректно выбирать порог на валидационной выборке, контролировать переобучение и отслеживать дрейф данных, т.к. при смене режимов работы системы качество реконструкции может снижаться.

2. Методы обнаружения контекстных аномалий.

2.1. Явное разделение контекстов.

Наиболее интерпретируемый подход – разделять данные по контекстуальным признакам (сезон, регион, класс объекта) и применять внутри каждого контекста методы точечных аномалий. Метод эффективен, если контексты заданы явно и внутри каждого контекста достаточно данных для стабильной оценки нормы [16]. Недостаток проявляется при большом числе контекстов и при плавных переходах между режимами: границы контекста становятся условными, а выбор корректного разделения может определять результат.

2.2. Модели остатков: регрессия и модели временных рядов.

Контекстные аномалии возможно выявлять через анализ остатков. В простейшем случае строится модель, предсказывающая значение x_b на основе контекстных признаков: $x_c: \hat{x}_b = f(x_c)$. Степень аномальности рассматриваемого объекта оценивается исходя из значения разности r между фактическими данными, характеризующими объект, и предсказанными моделью, т.е. $r = x_b - f(x_c)$. Чем больше значение r, тем больше оснований полагать, что рассматриваемый объект является аномальным относительно всех объектов в рассматриваемом наборе данных. При поиске аномалий во временных рядах этот подход реализуется с помощью методов прогнозирования временных рядов: модель предсказывает следующее значение временного ряда, а отклонение фактического значения от прогноза является основанием полагать, что рассматриваемый шаг временного ряда является аномалией. К классическим статистическим моделям для анализа и прогнозирования временных рядов относятся ARIMA-подобные модели. В дополнение к предыдущим подходам, в системах мониторинга часто применяют адаптивные пороги и алгоритмы обнаружения смены режима, в целях повышения качества работы моделей [17]. Преимуществом указанных подходов является объяснимость: модель показывает, какое значение ожидалось, а остаток r дает понять, насколько реальное наблюдение отклоняется от этого ожидания.

2.3. Рекуррентные нейросети и стохастические модели.

Архитектуры класса рекуррентных нейронных сетей LSTM (Long Short-Term Memory) и GRU (Gated Recurrent Unit) применяют для решения задач, в которых данные представлены в виде последовательностей, с целью прогнозирования, реконструкции данных, обучения векторных представлений (извлечения признаков) и классификации. При использовании рекуррентных нейронных сетей в задаче поиска аномалий в данных оценкой аномальности служит ошибка прогноза или реконструкции, в следствие чего такой подход естественным образом учитывает контекст предшествующих состояний. Для многомерных временных рядов используют стохастические рекуррентные модели. Они более устойчивы к шуму, пропускам в данных и неполноте разметки [11, 18]. Практический компромисс состоит в том, что такие модели требуют значительного объема данных, процедуры временной валидации и контроля переобучения, иначе модель начинает переобучаться.

3. Методы обнаружения коллективных аномалий.

3.1. Анализ окон и последовательностей.

Коллективные аномалии удобно искать на уровне окон фиксированной длины W, вычисляя статистики окна (среднее, дисперсия, спектральные признаки, частоты событий) и сравнивая их с распределением статистик для нормальных окон. Это простой и управляемый по ложным тревогам подход, который хорошо работает для скачков уровня и длительных дрейфов [18, 19]. Ограничение – зависимость от выбора W: слишком короткое окно не видит паттерн, слишком длинное размывает событие и затрудняет локализацию. На практике используют несколько окон и берут максимум оценки аномальности, либо применяют многошкальные признаки.

3.2. Кластеризация и плотностная сегментация.

Для табличных и пространственных данных коллективные аномалии проявляются как небольшие или разреженные кластеры, либо как группы точек, не принадлежащие ни одной плотной области. Алгоритм кластеризации K-means полезен как быстрый инструмент предварительного анализа данных, но требует выбора числа кластеров и плохо работает с кластерами произвольной формы (например, эллипсоидные или дугообразные). Алгоритм кластеризации, основанный на плотности данных DBSCAN (Density-Based Spatial Clustering of Applications with Noise) не требует заранее задавать число, характеризующее количество кластеров, и естественно выделяет шум и группы нетипичного поведения [7, 10]. В прикладных задачах поиска аномалий в данных кластеризацию применяют, чтобы объединить подозрительные объекты в группы, которые затем проверяют другими методами выявления аномалий. Такой подход упрощает анализ, поскольку рассматривать совокупность взаимосвязанных отклонений проще, чем множество изолированных объектов.

3.3. Последовательные автоэнкодеры и глубокие одноклассовые методы.

Для выявления аномалий в данных, которые характеризуют выполнение сложных паттернов, применяют автоэнкодеры (включая такие, как LSTM-автоэнкодер или GRU-автоэнкодер), работающие с последовательностями, а также глубокие одноклассовые методы (OC-NN, Deep SVDD). Автоэнкодеры ищут аномалии по ошибке восстановления, одноклассовые методы – по удалённости от компактной области нормы. Такие модели особенно эффективны при мониторинге технических систем, где коллективная аномалия проявляется как устойчивое изменение динамики процесса, а не как единичный выброс [11, 15]. Однако на практике важно предусмотреть интерпретируемость результатов и процедуры верификации. Без этого модель может выдавать сигналы тревоги, не позволяя понять их причины.

4. Выбор порога и калибровка.

Порог принятия решения T не является внешним гиперпараметром, т.к. он непосредственно влияет на баланс между полнотой и точностью. На практике применяют квантильные правила (например, считать аномалией верхние q процентов значений $s(x)$), методы экстремальных значений для потоковых данных, а также динамические пороги, зависящие от времени суток, сезонности или текущего режима работы системы. Выбор стратегии определяется стоимостью ошибок: в задачах кибербезопасности избыток ложных срабатываний перегружает операторов, а при мониторинге критического оборудования пропуск события недопустим. Поэтому настройка порога обычно требует согласования с предметной областью и расчёта ожидаемых издержек [8, 11].

5. Оценка качества и протоколы эксперимента.

Для бинарной классификации традиционно используют точность (precision), полноту (recall) и F-меру (F_1). Однако при сильном дисбалансе классов более надёжными считаются метрики PR-AUC (площадь под кривой точность–полнота) и коэффициент корреляции Мэттьюса (MCC). Для временных рядов важна не только поточечная точность, но и правильность детектирования событий в целом, т.е. раннее предупреждение, минимальная задержка, устойчивость на интервале. В прикладных задачах применяют событийные метрики и процедуры сглаживания разметки внутри одного события. Корректный протокол валидации должен исключать утечку информации: тестовый период должен быть отделён от обучающего по времени, а все преобразования (нормализация, кодирование) выполняться только на основе обучающей выборки [3].

РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЯ И ИХ ОБСУЖДЕНИЕ

Предложенная классификация типов аномалий и методов их обнаружения позволяет систематизировать подходы к построению систем мониторинга. Для наглядного сопоставления ключевых характеристик в таблице 1 сведены три основных типа аномалий: глобальные, контекстные и коллективные. По каждому типу указаны масштаб проявления, зависимость от контекста, типичные виды данных, достаточная информация для детектирования, примеры алгоритмов, а также степень интерпретируемости получаемых результатов.

Из таблицы 1 видно, что выбор метода в значительной степени определяется тем, как именно проявляется аномалия: является ли она изолированным выбросом, зависит ли от внешних условий или представляет собой групповой паттерн. Эти различия напрямую влияют на применимость статистических, плотностных, модельных и нейросетевых подходов.

В таблице 2 выполнено сопоставление классов методов обнаружения аномалий. Для каждого класса приведены типичные алгоритмы, наиболее подходящие типы аномалий, а также основные преимущества и ограничения. Такое сопоставление позволяет оценить, какие методы целесообразно применять на начальных этапах построения системы (например, статистические или изоляционные), а какие требуют более сложной настройки и значительных вычислительных ресурсов (кластеризация, глубокие модели).

Таблица 1

Сравнение типов аномалий

Table 1

Comparison of anomaly types

Параметр	Глобальная аномалия	Контекстная аномалия	Коллективная аномалия
Масштаб проявления	Одна точка	Одна точка	Группа точек, окно, последовательность
Зависимость от контекста	Нет	Да	Частично
Аномальность отдельного значения	Наиболее вероятно	Нет	Наименее вероятно
Типичные данные	Табличные признаки, сенсоры	Временные и пространственные данные	Временные ряды, логи, события
Достаточная информация для детектирования	Глобальная статистика	Контекстные атрибуты	Структура группы и взаимосвязи
Примеры методов	IQR, Mahalanobis, IF	Регрессия остатков, LSTM	Окна, DBSCAN, seq-autoencoder
Интерпретируемость	Высокая	Средняя	Низкая

Таблица 2

Классы методов и применимость

Table 2

Method families and applicability

Класс методов	Примеры алгоритмов	Наиболее подходящие типы аномалий	Ключевые преимущества и ограничения
Статистические	Z-score, IQR, MAD	Глобальные	Высокая скорость и интерпретируемость; ограничено предположениями о распределении и одномерными данными
Плотность/расстояния	kNN, LOF, Mahalanobis	Глобальные, локальные	Гибкость без параметрической модели; требует метрики расстояния, квадратичная сложность

Изоляция/границы	Isolation Forest, One-Class SVM	Глобальные, частично контекстные	Эффективно для табличных данных; ограниченная интерпретируемость, требует подбора гиперпараметров
Кластеризация	K-means, DBSCAN	Коллективные (кластеры)	Находит группы и шум; чувствительно к параметрам и масштабу признаков
Последовательные модели	ARIMA, HMM, change-point	Контекстные, коллективные	Учитывает динамику и режимы; требует корректной валидации во времени
Глубокая реконструкция	AE, VAE, seq-AE, DAGMM	Все типы (при корректной постановке)	Высокая выразительность; нужен объём данных и контроль наличие аномалий в обучающих данных

Построение системы обнаружения аномалий целесообразно вести поэтапно, от простых методов к более сложным. Базовые подходы (статистические, расстояния, изоляция) применяются в первую очередь. Если они не дают приемлемого результата, добавляется учёт контекста или анализ групповой структуры данных. Такая каскадная схема снижает вычислительные затраты и упрощает контроль качества.

Этапы выбора метода обнаружения аномалий:

1) Описать ожидаемую аномалию: точка; условная точка; группа.

2) Проверить наличие явного контекста (время, место, режим, класс).

3) Для глобальных аномалий: применить робастную статистику и простой ML-базис (IQR, Mahalanobis, LOF, Isolation Forest).

4) Для контекстных: добавить условную модель (стратификация, регрессия по контексту, модели временных рядов, LSTM).

5) Для коллективных: перейти к анализу окон, кластеризации и последовательных моделей (DBSCAN, окна, LSTM-автоэнкодер).

6) Настроить порог под цену ошибок и обеспечить мониторинг ложных тревог (квантили, EVT, динамические пороги).

7) Оценить качество на независимом временном отрезке (PR-AUC, MCC, событийные метрики).

Типичные ошибки, снижающие эффективность внедрения, можно свести к четырем группам.

1. Игнорирование контекста: одна и та же величина может быть нормальной и аномальной в разных режимах. В результате модель либо принимает за аномалию нормальные события, события, являющиеся угрозами.

2. Некорректная валидация: смешивание будущих наблюдений в обучении приводит к завышенным метрикам и к низким метрикам после ввода в эксплуатацию.

3. Наличие аномалий в обучающих данных: если модель обучается на «норме», но в тренировочном наборе много неразмеченных аномалий, автоэнкодер и другие модельные методы могут начать их воспроизводить, снижая чувствительность.

4. Отсутствие интерпретации: технические и организационные мероприятия требуют объяснений. Поэтому даже при использовании сложных нейросетевых моделей полезно сопровождать решение интерпретируемыми признаками, протоколом ручной проверки и механизмами локального анализа вклада признаков.

ВЫВОДЫ

В работе рассмотрены три базовых типа аномалий: глобальные, контекстные и коллективные, каждый из которых требует разных предположений о данных и разных классов методов. Глобальные аномалии часто эффективно обнаруживаются робастной статистикой, расстояниями, локальной плотностью и методами изоляции. Контекстные аномалии требуют условного анализа, который реализуется через стратификацию, модели остатков и последовательные нейросети. Коллективные аномалии предполагают анализ групп и паттернов, где применимы методы оконного анализа, кластеризации и последовательного представления данных.

Ключевой практический вывод состоит в том, что универсальный алгоритм выявления аномалий в данных отсутствует: надежные решения строятся как каскад (базовые проверки, применяемые совместно со сложные модели для выявленных отклонений от нормы) и как ансамбль (несколько независимых оценок аномальности) с последующей экспертной интерпретацией. Правильная типизация аномалии и корректная оценка качества (с учетом дисбаланса классов и природы происхождения данных) являются необходимыми условиями, чтобы методы машинного обучения и нейросетевые подходы работали на практике.

Список литературы

1. Ларин Д.О. Информационные революции и их роль в развитии человечества / Д.О. Ларин // Вестник Омского университета. – 2025. – Т. 30, № 1. – С. 37-50. – DOI 10.24147/1812-3996.2025.1.37-50. – EDN GJVYOV.

2. Шкодырев В.П. Обзор методов обнаружения аномалий в потоках данных / В.П. Шкодырев, К.И. Ягафаров, В.А. Баштовенко, Е.Э. Ильина // Труды Второй конференции по разработке программного обеспечения и информационному менеджменту (SEIM-2017). – СПб.: СПбГУТ, 2017. – Т. 1864. – С. 215-225.

3. Шоргин С.Я. Статистика и кластеры в поисках аномальных вкраплений в условиях больших данных // Информатика и её применения. – 2021. – Т. 15, № 4. – С. 142–151. – DOI: 10.15393/j12.art.2021.7987.

4. Андрианова Е.Г., Головин С.А., Зыков С.В., Лесько С.А., Чукалина Е.Р. Обзор современных моделей и методов анализа временных рядов динамики процессов в социальных, экономических и социотехнических системах // Российский технологический журнал. – 2020. – Т. 8, № 4. – С. 7-45. – DOI: 10.32362/2500-316X-2020-8-4-7-45.

5. Видищева Е.В., Копырин А.С., Василенко М.С. Анализ и уточнение классификации аномалий и выбросов на экономических данных // Вестник Алтайской академии экономики и права. – 2019. – № 6-1. – С. 41–46. – URL: https://vaael.ru/ru/article/view?id=589 (дата обращения: 22.01.2026).

6. Андрианова Е.Г., Зыков С.В. и др. Обзор современных моделей и методов анализа временных рядов // Российский технологический журнал. – 2020. – Т. 8, № 4. – С. 7-45.

7. Бардасова И.А., Волкова Е.А. Обнаружение аномалий в электронных письмах с помощью машинного обучения // Вестник науки №5(74). – Т. 4. – С. 1350-1358. 2024 г. // URL: https://www.вестник-науки.рф/article/14991 (дата обращения: 22.01.2026 г.).

8. Михайлов А.Н. Обнаружение аномалий в сетевом трафике с использованием методов машинного обучения // Вестник науки №12(81). – Т. 3. – С. 1463-1466. 2024 г. // URL: https://www.вестник-науки.рф/article/19907 (дата обращения: 22.01.2026 г.).

9. Домашкин А.А. Применение двухэтапного метода кластеризации для обнаружения аномалий: тез. докл. / А.А. Домашкин // Международная конференция по компьютерным системам и технологиям (ICCSS-2024). – М.: ИПУ, 2024. – С. 112-120. – URL: https://iccss2024.ipu.ru/proceedings/Домашкин.pdf (дата обращения: 22.01.2026).

10. Глухов К.А. Применение двухэтапного метода кластеризации / К.А. Глухов, А.А. Домашкин // Безопасные и информационные технологии. – 2023. – Т. 1, № 1. – С. 1-10. – URL: https://info-secur.ru/index.php/ojs/article/view/482 (дата обращения: 22.01.2026).

11. Краева Я.А. Нейросетевой метод обнаружения аномалий в многомерных потоковых временных рядах // Вестник СПбПУ. Сер. Радиотехника, телекоммуникации и средств вычислительной техники. – 2024. – № 2. – С. 45-58.

12. Гриценко А.В. Типы аномалий в видеоизображениях // Прикладная информатика. – 2012. – № 5. – С. 78-92.

13. Литвинович А.В., Смирнов С.В. Методы анализа многомерных данных в задачах обнаружения аномалий // Программные продукты и системы. – 2022. – Т. 135, № 3. – С. 45-52.

14. Герасимов М.А., Петров И.В. Выявление аномалий в масштабных данных с применением Isolation Forest и автоэнкодера // Вестник СПбГУ. Серия 15. Вычислительная математика и информатика. – 2024. – Т. 20, № 1. – С. 112—125.

15. Левшун Д.А., Попов Д.А., Козлов А.С. Обнаружение и объяснение аномалий в промышленных системах IoT на основе автоэнкодеров // Программные продукты и системы. – 2023. – Т. 141, № 4. – С. 123-135.

16. Бутусов Д.Н. Численные методы анализа нестационарных сигналов в задачах обработки изображений: дис. канд. физ.-мат. наук: 05.12.04 / Бутусов Д.Н.; СПбГЭТУ «ЛЭТИ». – СПб., 2021. – 150 с.

17. Носко В. П. Введение в регрессионный анализ временных рядов // [учебное пособие]. – М.: ВШЭ, 2010. – 120 с.

18. Брыкин Д.О. Исследование алгоритмов обработки временных рядов с учетом нестационарности: дис. канд. физ.-мат. наук: 05.13.18 / Брыкин Д.О.; МФТИ. – М., 2023. – 145 с.

19. Любушин А. А. Анализ данных систем геофизического и инженерного мониторинга. – 3-е изд. – М.: Наука, 2024. – 320 с.

References

1. Larin D.O. Information revolutions and their role in the development of mankind / D.O. Larin // Bulletin of Omsk University. – 2025. - Vol. 30, No. 1. – Pp. 37-50. – DOI 10.24147/1812-3996.2025.1.37-50. – EDN GJVYOV.

2. Shkodyrev V.P. Overview of anomaly detection methods in data streams / V.P. Shkodyrev, K.I. Yagafarov, V.A. Bashtovenko, E.E. Ilyina // Proceedings of the Second Conference on Software Engineering and Information Management (SEIM-2017). – St. Petersburg: SPbSUT, 2017. – Vol. 1864. – Pp. 215-225.

3. Shorgin S.Ya. Statistics and clusters in search of anomalous inclusions under big data conditions // Informatics and Its Applications. – 2021. – Vol. 15, No. 4. – Pp. 142-151. – DOI: 10.15393/j12.art.2021.7987.

4. Andrianova E.G., Golovin S.A., Zykov S.V., Les'ko S.A., Chukalina E.R. Review of modern models and methods for analyzing time series dynamics in social, economic, and sociotechnical systems // Russian Technological Journal. – 2020. – Vol. 8, No. 4. – Pp. 7-45. – DOI: 10.32362/2500-316X-2020-8-4-7-45.

5. Vidishcheva E.V., Kopyrin A.S., Vasilenko M.S. Analysis and refinement of anomaly and outlier classification on economic data // Bulletin of the Altai Academy of Economics and Law. – 2019. – No. 6-1. – Pp. 41-46. – URL: https://vaael.ru/ru/article/view?id=589 (date of access: 22.01.2026).

6. Andrianova E.G., Zykov S.V., et al. Review of modern models and methods for time series analysis // Russian Technological Journal. – 2020. – Vol. 8, No. 4. – Pp. 7-45.

7. Bardasova I.A., Volkova E.A. Anomaly detection in emails using machine learning // Bulletin of Science, No. 5(74), Vol. 4, pp. 1350-1358. 2024. ISSN 2712-8849 // URL: https://www.vesnik-nauki.rf/article/14991 (Accessed: 22.01.2026).

8. Mikhailov A.N. Anomaly detection in network traffic using machine learning methods // Bulletin of Science, No. 12(81), Vol. 3, pp. 1463-1466. 2024. ISSN 2712-8849 // URL: https://www.vesnik-nauki.rf/article/19907 (Accessed: 22.01.2026).

9. Domashkin A.A. Application of two-stage clustering method for anomaly detection: Conference abstract / A.A. Domashkin // International Conference on Computer Systems and Technologies (ICCSS-2024). – Moscow: IPU, 2024. – Pp. 112-120. – URL: https://iccss2024.ipu.ru/proceedings/Домашкин.pdf (accessed: 22.01.2026).

10. Glukhov K.A. Application of two-stage clustering method based on self-organizing Kohonen map for anomaly detection in synthetic datasets / K.A. Glukhov, A.A. Domashkin // Secure and Information Technologies. – 2023. – Vol. 1, No. 1. – Pp. 1-10. – URL: https://info-secur.ru/index.php/ojs/article/view/482 (date of access: 22.01.2026).

11. Kraeva Ya.A. Neural network method for anomaly detection in multidimensional streaming time series // Bulletin of SPbPU. Series: Radio Engineering, Telecommunications and Computer Engineering. – 2024. – No. 2. – Pp. 45-58.

12. Gritsenko A.V. Types of anomalies in video images // Applied Informatics. – 2012. – No. 5. – Pp. 78-92.

13. Litvinovich A.V., Smirnov S.V. Methods for analyzing multidimensional data in anomaly detection tasks // Software Products and Systems. – 2022. – Vol. 135, No. 3. – P. 45-52.

14. Gerasimov M.A., Petrov I.V. Anomaly Detection in Large-Scale Data Using Isolation Forest and an Autoencoder // Bulletin of St. Petersburg State University. Series 15. Computational Mathematics and Informatics. – 2024. – Vol. 20, No. 1. – P. 112–125.

15. Levshun D.A., Popov D.A., Kozlov A.S. Detection and explanation of anomalies in industrial IoT systems based on autoencoders // Software Products and Systems. – 2023. – Vol. 141, No. 4. – P. 123–135.

16. Butusov D.N. Numerical methods for analyzing non-stationary signals in image processing tasks tasks [Doctoral dissertation, Cand. Phys.-Math. Sci., 05.12.04]. SPbGETU "LETİ". – St. Petersburg, 2021. – 150 p.

17. Nosko V.P. Introduction to regression analysis of time series // [study guide]. – Moscow: HSE, 2010. – 120 p.

18. Brykin D.O. Investigation of time series processing algorithms considering non-stationarity [Doctoral dissertation, Cand. Phys.-Math. Sci., 05.13.18]. MIPT. – Moscow, 2023. – 145 p.

19. Lyubushin A.A. Analysis of geophysical and engineering monitoring system data. – 3rd ed. – M.: Nauka, 2024. – 320 p.

Котов Дмитрий Васильевич, доктор технических наук, доцент, начальник 3 центра научно-исследовательского института (информационной безопасности), Военная академия Генерального штаба Вооружённых сил Российской Федерации, г. Москва, Россия

Kotov Dmitry Vasilyevich, Doctor of Technical Sciences, Associate Professor, Head of the 3rd Center of the Scientific Research Institute (Information Security), Military Academy of the General Staff of the Armed Forces of the Russian Federation, Moscow, Russia