

УДК 004.032.26

DOI: 10.24412/3033-6007-2026-339-53-68

СРАВНИТЕЛЬНОЕ ИССЛЕДОВАНИЕ МЕТОДОВ ОБНАРУЖЕНИЯ ВЫБРОСОВ В ДАННЫХ МОНИТОРИНГА НЕФТЯНЫХ СКВАЖИН НА ОСНОВЕ SVM И DBSCAN*

Михайлов И.С.^{1, а}, Мьо Хлайн Вин^{1, b}, Сидоров К.О.^{1, c}

¹ Национальный исследовательский университет «Московский энергетический институт» (НИУ МЭИ), Москва, Россия.

E-mail: ^a fr82@mail.ru, ^b myohlaingwin69287@gmail.com, ^c kirill.sidoroff2014@yandex.ru

Аннотация. Цель исследования – сопоставить результаты One-Class Support Vector Machine (One-Class SVM) и DBSCAN при обнаружении статистических аномалий в данных мониторинга нефтяной скважины и оценить возможность их совместного использования. В анализе использованы шесть параметров многофазного измерителя; после преобразования методом главных компонент (PCA) они представлены тремя главными компонентами, сохраняющими 89,86% общей дисперсии. На выборке из 8639 наблюдений One-Class SVM выявил 424 аномальные точки, DBSCAN – 188; 85 наблюдений обнаружены обоими методами. Для классификации режимов течения по уровню газосодержания SVM с RBF-ядром ($C = 100,0$, $\gamma = 1,0$) на стратифицированном разбиении 70/30 показал точность 98,26% при F1-мере 0,97–0,99. При 5-кратной стратифицированной кросс-валидации F1_масго составила $0,9853 \pm 0,0026$. Совместное рассмотрение результатов позволяет выделять согласованные двумя методами, обнаруженные только SVM и обнаруженные только DBSCAN статистические выбросы для последующей экспертной проверки.

Ключевые слова: обнаружение выбросов, метод опорных векторов, DBSCAN, метод главных компонент, мониторинг нефтяных скважин, машинное обучение, классификация режимов течения, аномалии.

Введение

Цель настоящего исследования – сравнительный анализ результатов двух методов обнаружения аномалий (One-Class SVM и DBSCAN) на данных мониторинга нефтяной скважины и разработка практических рекомендаций по их совместному применению. Совместные и отдельные срабатывания используются для ранжирования статистических выбросов по приоритету экспертной проверки.

Для достижения цели поставлены следующие задачи:

1. сформировать признаковое пространство из шести параметров, используемых в дальнейшем анализе;
2. применить метод главных компонент для снижения размерности и определить число компонент, сохраняющих 89,86% общей дисперсии;

Для цитирования:

Михайлов И.С., Мьо Хлайн Вин, Сидоров К.О. Сравнительное исследование методов обнаружения выбросов в данных мониторинга нефтяных скважин на основе SVM и DBSCAN. *Интеллектуальный транспорт*. 2026. № 3(39). С. 53–68. DOI: 10.24412/3033-6007-2026-339-53-68.

* Исследование было проведено при поддержке РНФ (проект № 24-11-00285).

3. выполнить обнаружение статистических аномалий методами One-Class SVM и DBSCAN на преобразованных данных;
4. сопоставить результаты двух методов и разделить обнаруженные точки по сочетанию срабатываний алгоритмов;
5. обучить и оценить SVM-классификатор режимов течения по уровням газосодержания на разбиении 70/30 и при 5-кратной стратифицированной кросс-валидации;
6. сформулировать практические рекомендации по совместному применению методов в системах мониторинга.

Статья организована следующим образом. В разделе 1 представлена предметная область и мотивация исследования. В разделе 2 приведено описание исходных данных, в разделе 3 – применение PCA, в разделе 4 – построение и оценка SVM-классификатора режимов течения, в разделе 5 – обнаружение локальных аномалий методом DBSCAN. Раздел 6 посвящён сравнению One-Class SVM и DBSCAN, раздел 7 – классификации результатов по сочетанию срабатываний методов, раздел 8 – практическим рекомендациям, раздел 9 – выводам по экспериментальному исследованию. В заключении обобщены результаты и обозначены направления дальнейшей работы.

Современные системы мониторинга нефтяных скважин регистрируют множество взаимосвязанных параметров, однако традиционный пороговый контроль рассматривает их раздельно. В результате сложные сочетания признаков могут быть пропущены, а отклонения отдельных показателей – приводить к ложным срабатываниям. Дополнительное ограничение связано с дефицитом размеченных данных о нештатных ситуациях, поэтому для контроля представляют интерес методы машинного обучения без учителя, способные работать с многомерными потоками данных и интегрироваться в системы управления [1,2].

Подходы к обнаружению аномалий включают статистические, дистанционные, плотностные, опорно-векторные и древовидные методы. One-Class SVM формирует глобальную границу нормальных наблюдений, тогда как DBSCAN выявляет локальные точки в областях пониженной плотности, поэтому их совместное применение позволяет анализировать разные типы отклонений. Для устранения корреляционной избыточности и подготовки данных к визуализации используется PCA. Сравнительное исследование One-Class SVM и DBSCAN именно на данных мониторинга нефтяных скважин и методика интерпретации пересечения результатов представлены в литературе неполно [3,4].

Сравнение Isolation Forest и One-Class SVM при обнаружении аномалий в данных эксплуатации газопроводов представлено в работе [5]. Применение методов обнаружения выбросов без учителя для восстановления каротажных данных в осложнённых условиях бурения рассмотрено в работе [6].

Нейросетевые модели с объяснением вклада признаков методом SHAP [7] рассматриваются как альтернатива для дальнейшего сравнения; в эксперимент они не включались.

Исследование направлено на сопоставление глобального и локального методов обнаружения аномалий после снижения размерности и приоритизацию экспертной проверки выбросов при дефиците размеченных данных.

1. Предметная область

Пороговый контроль показаний датчиков нефтяных скважин не учитывает взаимосвязи параметров, что может приводить к ложным срабатываниям и пропуску аномалий [1]. Сложность получения размеченных данных о нештатных ситуациях обуславливает интерес к методам машинного обучения без учителя для автоматического выявления аномалий в режиме реального времени.

С практической стороны в данном исследовании рассматривается обнаружение статистических аномалий в показаниях многофазного измерителя обводнённости, установленного на устье нефтяной скважины. Данный измеритель регистрирует в режиме реального времени набор параметров, позволяющий оценивать профиль течения нефте-водо-газового потока и отслеживать его изменения.

2. Исходные данные

Многофазный измеритель обводнённости нефтяных скважин устанавливается на вертикальном участке нефтепровода. Измеритель оснащается двумя ультразвуковыми сенсорами, расположенными в центре и на периферии нефте-водо-газового потока. Измеритель регистрирует значения скорости ультразвука в центре и на периферии потока, газосодержание в центре и на периферии потока, температуру потока, обводнённость в центре и на периферии потока. На основании расчётов измеритель вычисляет интегральное значение обводнённости и газосодержания потока. Данные записываются в архив измерителя один раз в десять секунд.

Процедура предобработки включала:

1. удаление дублирующей строки заголовка;
2. приведение числовых столбцов к типу float с использованием `pd.to_numeric(errors='coerce')`;
3. удаление строк с полностью отсутствующими значениями;
4. удаление строк с пропусками в выбранных признаках.

Интегральное газосодержание используется для формирования целевой переменной и не включается во входное признаковое пространство. Классы низкого, среднего и высокого газосодержания задаются порогами 0,3 и 0,7. Их объёмы составляют соответственно 2184, 2602 и 3853 наблюдения. В анализируемую выборку вошли 8639 наблюдений. Это соответствует сумме категорий, представленных на рисунках 4 и 7. Исследование выполнено на данных многофазного измерителя. Переносимость результатов на другие скважины и оборудование требует отдельной проверки.

Удаление пропусков само по себе не позволяет отличить сбой датчика от технологического отклонения или дрейфа калибровки. В настоящем исследовании эти причины статистических выбросов не разграничиваются.

3. Метод главных компонент (PCA)

Метод главных компонент (Principal Component Analysis, PCA) применяется в данном исследовании для снижения размерности данных мониторинга нефтяных скважин. Исходные данные содержат шесть взаимосвязанных параметров (скорость звука в центре и на периферии, газосодержание в центре и на периферии, температура, интегральная обводнённость), многие из которых коррелируют между собой, что создаёт избыточность информации и затрудняет визуализацию [3, 4]. PCA позволяет преобразовать исходные признаки в новые переменные – главные компоненты, которые являются линейными комбинациями исходных признаков и упорядочены по убыванию дисперсии.

Для определения числа главных компонент применяется оценка объяснённой и кумулятивной дисперсии. Перед применением PCA каждый признак стандартизируется:

$$x_{\text{scaled}} = \frac{x - \mu}{\sigma}, \quad (1)$$

где x – исходное значение признака; μ – его среднее значение; σ – стандартное отклонение.

Первая главная компонента (PC1) объясняет 52,15% общей дисперсии и отражает преимущественно газосодержание и температуру потока: положительные нагрузки для газосодержания в центре (0,510), температуры (0,460) и газосодержания на периферии (0,248) сочетаются с отрицательными нагрузками для скорости звука в центре (–0,490) и интегральной обводнённости (–0,461). Вторая компонента (PC2) объясняет 27,51% дисперсии и связана с распределением параметров по сечению потока (нагрузки Звук_П = 0,706, Газ_П = 0,481, Вода_Инт = 0,409). Третья компонента (PC3) объясняет 10,20% дисперсии и отражает преимущественно периферийное газосодержание (Газ_П = 0,831) с отрицательными нагрузками температуры (–0,371) и скорости звука на периферии (–0,350). Совокупно первые три компоненты сохраняют 89,86% исходной информации, что позволяет использовать их для дальнейшего анализа вместо исходных шести признаков [3].

Эти три компоненты используются как входные признаки последующих моделей. Рисунки 2–4 показывают двумерную проекцию PC1-PC2 для наглядного сопоставления результатов, тогда как рисунки 5–7 используют трёхмерное пространство PC1-PC2-PC3 и тем самым учитывают все три выбранные компоненты.

4. Классификация с помощью метода опорных векторов SVM

Метод опорных векторов (Support Vector Machine, SVM) применяется в данном исследовании для классификации режимов течения по уровню газосодержания. В отличие от One-Class SVM, используемого для обнаружения аномалий, классический SVM решает задачу многоклассовой классификации, разделяя данные на три класса: низкое, среднее и высокое газосодержание. Результаты PCA и доли объяснённой дисперсии для главных компонент представлены в таблице 1.

Входными признаками для классификатора служат три главные компоненты (PC1, PC2, PC3), полученные в результате PCA-преобразования.

Таблица 1. Дисперсия, объясняемая главными компонентами

Главная компонента	Объясненная дисперсия (%)	Кумулятивная дисперсия (%)
PC1	52,15	52,15
PC2	27,51	79,66
PC3	10,20	89,86
PC4	6,14	96,00
PC5	2,73	98,73
PC6	1,27	100,0

Использование ядра RBF (Radial Basis Function) позволяет строить нелинейные разделяющие границы. В представленном эксперименте использованы значения параметра регуляризации $C = 100,0$ и параметра RBF-ядра $\gamma = 1,0$. Формула 2 задаёт решающую функцию SVM, применяемую после обучения модели.

$$f(x) = \sum_{i=1}^n \alpha_i y_i K(x_i, x) + b, \quad (2)$$

где n – количество данных в обучающей выборке; α_i – множители Лагранжа, полученные во время обучения (ненулевые только для опорных векторов); y_i – истинная метка класса (± 1) для i -го наблюдения; x_i – обучающие данные; $K(x_i, x)$ – функция ядра (например, линейная, RBF), которая вычисляет сходство между x_i и новой точкой x без явного вычисления многомерных координат.

В формуле 2 значение $f(x)$ является решающим значением классификатора для новой точки x и вычисляется по опорным векторам, их коэффициентам и выбранной функции ядра. Параметр $C = 100,0$ используется на этапе обучения как параметр регуляризации; представленные метрики относятся к указанной конфигурации модели.

Обучение SVM производится на 70% данных, оставшиеся 30% используются для тестирования. Стратификация выборки обеспечивает сохранение пропорций классов в обучающей и тестовой выборках.

Оценка качества классификации выполнена на тестовой части выборки. Полученная точность составила 98,26%, а значения F1-меры для классов низкого, среднего и высокого газосодержания находились в диапазоне 0,97–0,99.

Полученные показатели относятся к одному стратифицированному разбиению 70/30 и к указанной конфигурации SVM с тремя главными компонентами и RBF-ядром. Дополнительная оценка выполнена с помощью 5-кратной стратифицированной кросс-валидации (табл. 2).

Первоначально использованы параметры $C = 10,0$, $\gamma = \text{auto}$, что обеспечило точность 98,42% на тестовой выборке.

Для обоснованного выбора гиперпараметров SVC выполнен сеточный поиск (Grid Search) с 5-кратной стратифицированной кросс-валидацией. Сетка параметров охватывала $C \in \{0,1, 1,0, 10,0, 100,0\}$ и $\gamma \in \{\text{scale}, \text{auto}, 0,01, 0,1, 1,0\}$ при фиксированном RBF-ядре. Целевая метрика оптимизации – F1_macro. Лучшая комбинация: $C = 100,0$, $\gamma = 1,0$, ядро RBF, обеспечившая F1_macro = 0,9853 на кросс-валидации.

Таблица 2. Результаты 5-кратной кросс-валидации SVM-классификатора

Метрика	Среднее	Стандартное отклонение
Accuracy	0,9867	0,0025
Precision_macro	0,9853	0,0024
Recall_macro	0,9854	0,0029
F1_macro	0,9853	0,0026

Средние значения и стандартные отклонения характеризуют вариацию метрик по пяти разбиениям. Выбор гиперпараметров выполнялся по F1_macro, поэтому он не равнозначен максимизации точности на тестовой части. Эти оценки не заменяют матрицу ошибок и независимую проверку на новых периодах наблюдений. Для проверки отсутствия утечки данных необходимо также фиксировать, что масштабирование признаков и PCA обучаются только на обучающей части каждого разбиения.

5. Метод DBSCAN

Для обнаружения локальных аномалий в данных мониторинга нефтяных скважин применён алгоритм пространственной кластеризации на основе плотности DBSCAN. Основное преимущество данного метода заключается в способности выявлять кластеры произвольной формы и отсутствии необходимости заранее задавать количество кластеров. Ключевыми параметрами метода являются радиус окрестности ε и минимальное количество точек $minPts$ [8]. В ходе исследования использованы значения $\varepsilon = 0,5$ и $minPts = 5$. Входными данными служат три главные компоненты, полученные в результате преобразования методом главных компонент.

В результате применения алгоритма DBSCAN выявлено 188 аномальных наблюдений, что составляет 2,18% от общего объёма данных. Данный результат показывает, что DBSCAN позволяет выявлять локальные выбросы, физическая природа которых требует проверки по исходным измерениям и эксплуатационным данным. В отличие от метода опорных векторов, который ориентирован на глобальные аномалии, DBSCAN дополняет статистическое описание данных, обнаруживая точки, находящиеся в областях с низкой плотностью данных [9]. Само по себе число срабатываний не подтверждает повышение надежности обнаружения реальных нарушений.

Для трёхмерного пространства PC1-PC3 выбор $minPts = 5$ удовлетворяет эвристике $minPts \geq dim + 1 = 4$. Построен график расстояний до k -го ближайшего соседа ($k = minPts - 1 = 4$) для точек в пространстве главных компонент (рис. 1). Точка максимальной кривизны, определяемая по второй производной функции k -расстояний, соответствует значению $\varepsilon \approx 16,17$.

Полученное значение не совпадает с использованным $\varepsilon = 0,5$, поэтому данный график не может служить подтверждением оптимальности $\varepsilon = 0,5$. Условие для $minPts$ также не определяет единственное оптимальное значение. Настройка $\varepsilon = 0,5$, $minPts = 5$ рассматривается как исследуемая конфигурация.

Для оценки устойчивости результатов к выбору гиперпараметров проведён анализ чувствительности. При варьировании ε от 0,3 до 1,0 число выявленных выбросов изменялось в диапазоне [24, 558], при этом число

кластеров уменьшалось с 39 до 4.

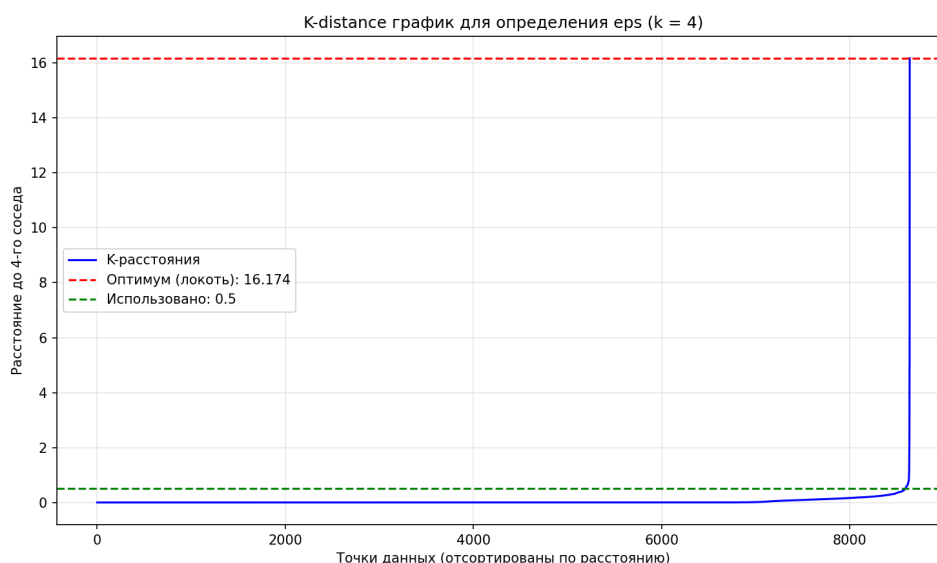


Рис. 1. График k -расстояний для DBSCAN ($k = 4$) и сравнение вычисленной точки максимальной кривизны с использованным $\varepsilon = 0,5$.

Таблица 3. Чувствительность DBSCAN к ε ($minPts = 5$)

ε	Выбросов	Доля, %	Кластеров
0,3	558	6,46	39
0,4	340	3,94	26
0,5	188	2,18	19
0,6	112	1,30	13
0,7	82	0,95	12
0,8	58	0,67	10
1,0	24	0,28	4

Таблица 4. Чувствительность DBSCAN к $minPts$ ($\varepsilon = 0,5$)

$minPts$	Выбросов	Доля, %	Кластеров
3	122	1,41	24
5	188	2,18	19
7	278	3,22	16
10	317	3,67	13
15	420	4,86	15
20	546	6,32	12

При уменьшении ε с 0,5 до 0,4 число выбросов возрастает со 188 до 340, а при увеличении до 0,6 сокращается до 112. Изменение $minPts$ с 5 до 3 или 7 приводит соответственно к 122 или 278 срабатываниям. Такая вариация не дает оснований считать результат устойчивым только на основе близости используемых настроек. При эксплуатации необходима проверка

качества срабатываний, а не только их числа.

6. Сравнение методов обнаружения аномалий SVM и DBSCAN

В ходе исследования проведён сравнительный анализ двух методов обнаружения аномалий – One-Class SVM и DBSCAN. Метод опорных векторов выявил 424 аномалии, тогда как DBSCAN обнаружил 188 аномалий. Пересечение результатов составило 85 образцов, или 16,13% от объединения срабатываний (527 уникальных наблюдений) [10]. Низкая степень пересечения (16,13%) допускает две интерпретации: методы обнаруживают различные типы аномалий; низкая согласованность может указывать на недостаточную надёжность хотя бы одного из методов в данной предметной области. Разделение срабатываний на категории не разрешает эту неоднозначность без независимой эксплуатационной валидации.

Для оценки устойчивости результатов One-Class SVM проведён анализ чувствительности к параметрам ν и γ . При варьировании $\nu \in \{0,01, 0,03, 0,05, 0,07, 0,10, 0,15\}$ число выбросов изменялось в диапазоне [88, 1233].

При $\nu = 0,05$ выявлено 424 наблюдения (4,91%). При этом значении ν в дополнительном анализе сравнены варианты $\gamma \in \{\text{scale}, \text{auto}, 0,001, 0,01, 0,1, 1,0\}$; для scale и auto получено по 424 срабатывания. При варьировании γ диапазон выбросов составил [424, 473], что свидетельствует о существенно меньшей чувствительности результатов к параметру γ по сравнению с ν .

Таблица 5. Чувствительность One-Class SVM к ν

ν	Выбросов	Доля, %
0,01	88	1,02
0,03	295	3,41
0,05	424	4,91
0,07	579	6,70
0,10	933	10,80
0,15	1233	14,27

Таблица 6. Чувствительность One-Class SVM к γ ($\nu = 0,05$)

γ	Выбросов	Доля, %
scale	424	4,91
auto	424	4,91
0,001	456	5,28
0,01	439	5,08
0,1	437	5,06
1,0	473	5,48

6.1. Сравнение с базовым пороговым методом

Выполнено сравнение с правилом 3σ (табл. 7). Пороговый метод классифицирует наблюдение как аномальное, если хотя бы один из шести признаков выходит за пределы $\mu \pm 3\sigma$.

Таблица 7. Сравнение методов обнаружения аномалий

Метод	Обнаружено	Доля, %
Пороговый (правило 3σ)	88	1,02
One-Class SVM	424	4,91
DBSCAN	188	2,18
Пересечение ($SVM \cap DBSCAN$)	85	0,98

Согласованность порогового метода с One-Class SVM составила 71 образец (16,7% от результатов SVM), с DBSCAN – 55 образцов (29,3% от результатов DBSCAN).

Сравнение показывает различия в числе и составе срабатываний, но не устанавливает превосходство одного метода над другим. Большее число отмеченных наблюдений может включать как дополнительные полезные обнаружения, так и ложные тревоги.

7. Классификация выбросов

На основе сочетания срабатываний двух методов предложена следующая классификация результатов (табл. 8).

Таблица 8. Классификация результатов по сочетанию SVM и DBSCAN

Результат совместной классификации	Количество	Критерий отнесения
Согласованные выбросы, приоритетная проверка	85	Выявлены обоими методами
Потенциальные аномалии, экспертная оценка	339	Выявлены только One-Class SVM
Локальные выбросы, отдельная проверка	103	Выявлены только DBSCAN

8. Практические рекомендации

Результаты работы методов представлены на рисунках 2–7. На основе полученных результатов предложена следующая стратегия использования методов:

1. **Пересечение** (85 образцов) – точки, выявленные обоими методами. Их следует рассматривать как приоритетные кандидаты для экспертной и, при наличии эксплуатационных данных, технической проверки, а не как автоматически подтверждённые аварийные состояния. Сопоставление результатов двух методов представлено на рисунке 4 в проекции PC1-PC2 и на рисунке 7 в пространстве PC1-PC2-PC3.
2. **Только SVM** (339 образцов) – точки, выявленные только One-Class SVM. Для них рекомендуется отдельная экспертная оценка и сопоставление

с временным контекстом мониторинга. Результаты One-Class SVM представлены на рисунке 2 в проекции PC1-PC2 и на рисунке 5 в пространстве PC1-PC2-PC3.

3. **Только DBSCAN** (103 образца) – точки, выявленные только DBSCAN. Их следует рассматривать как локальные статистические выбросы или кандидаты на шум и не исключать автоматически без дополнительной проверки. Результаты DBSCAN представлены на рисунке 3 в проекции PC1-PC2 и на рисунке 6 в пространстве PC1-PC2-PC3.

Для последующего анализа категории «Пересечение», «Только SVM» и «Только DBSCAN» следует сохранять отдельно. Без независимой эксплуатационной разметки эти категории отражают сочетание срабатываний алгоритмов. Физическая интерпретация точки требует сопоставления координат PC1-PC3 с соответствующими исходными показаниями и временем регистрации. Анализ скорости звука, температуры и газосодержания в центре и на периферии потока совместно с эксплуатационными журналами должен предшествовать выводу о неисправности оборудования. Отсутствие независимой событийной разметки не позволяет связывать 85 согласованных срабатываний с технологическими нарушениями.

9. Выводы по результатам экспериментального исследования

Эксперимент показал, что PCA сократил размерность с шести до трёх главных компонент при сохранении 89,86% общей дисперсии. На выборке из 8639 наблюдений One-Class SVM выявил 424 аномальные точки, DBSCAN – 188; общими для двух методов оказались 85 наблюдений, что составляет 16,13% от объединения срабатываний двух методов (527 уникальных точек).

SVM-классификация достигла точности 98,26% при F1-мере 0,97–0,99. В практическом мониторинге пересечение методов следует считать приоритетными выбросами, результаты «только SVM» направлять на экспертную проверку, а точки «только DBSCAN» рассматривать как локальные статистические отклонения, требующие отдельной проверки. При 5-кратной стратифицированной кросс-валидации получены $accuracy = 0,9867 \pm 0,0025$ и $F1_macro = 0,9853 \pm 0,0026$. Пороговый метод по правилу 3σ выявил 88 наблюдений (1,02%).

Заключение

Применение PCA позволило сократить размерность данных с 6 до 3 компонент с сохранением 89,86% информации. One-Class SVM выявил 424 аномалии, DBSCAN – 188, из них 85 обнаружены обоими методами [11]. Выделены три категории: согласованные выбросы – 85 наблюдений, только SVM – 339, только DBSCAN – 103. Классификация режимов течения на основе SVM показала точность 98,26% при F1-мере 0,97–0,99 для трёх классов.

Научная новизна состоит в совместном анализе One-Class SVM и DBSCAN на данных мониторинга нефтяной скважины и исследовании чувствительности результатов к гиперпараметрам. Практическая значимость связана с возможностью применения подхода в системах мониторинга, раннего предупреждения и прогнозного обслуживания.

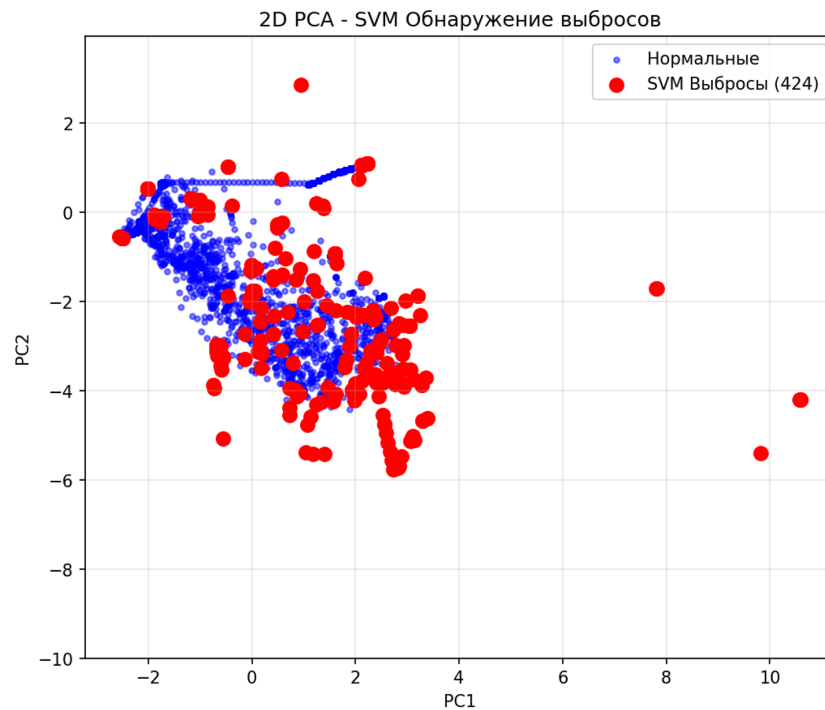


Рис. 2. Обнаружение выбросов методом One-Class SVM в пространстве PC1-PC2

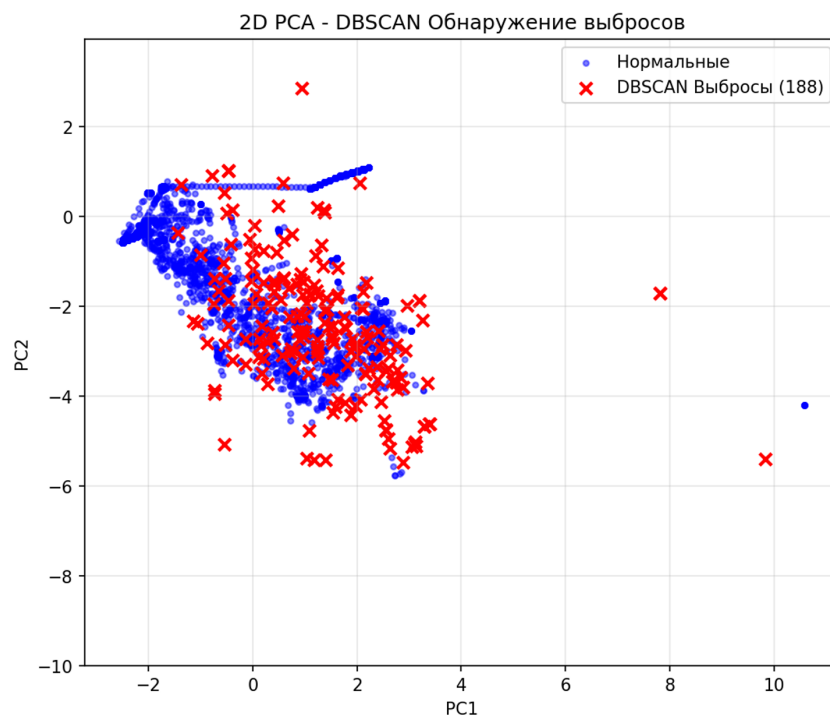


Рис. 3. Обнаружение выбросов методом DBSCAN в пространстве PC1-PC2

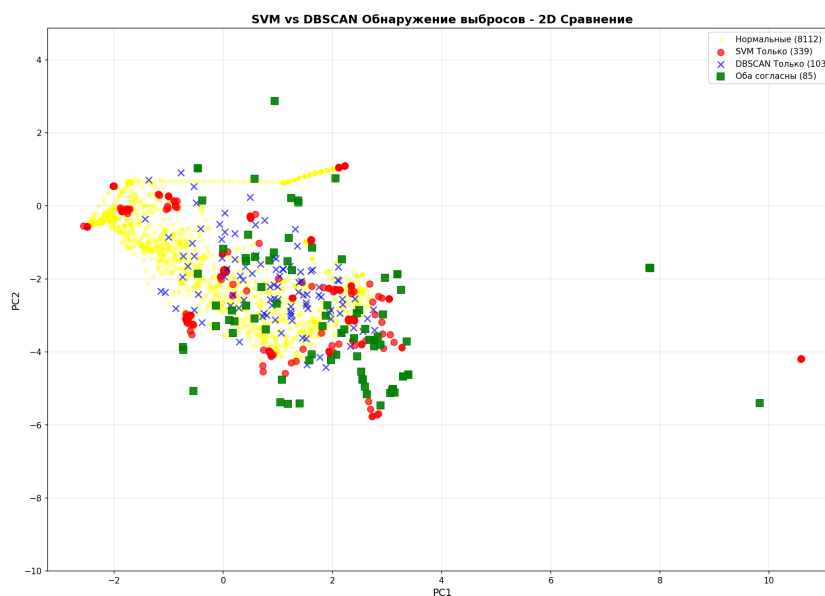


Рис. 4. Сравнение результатов обнаружения выбросов методами SVM и DBSCAN в пространстве PC1-PC2

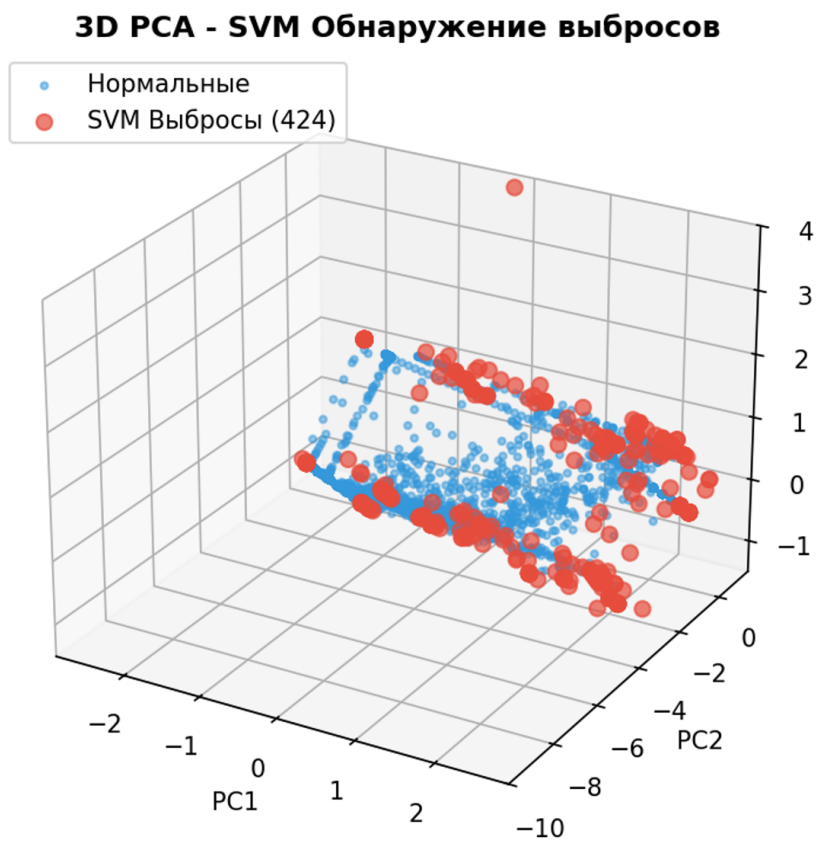


Рис. 5. Трёхмерная визуализация аномалий, обнаруженных методом One-Class SVM

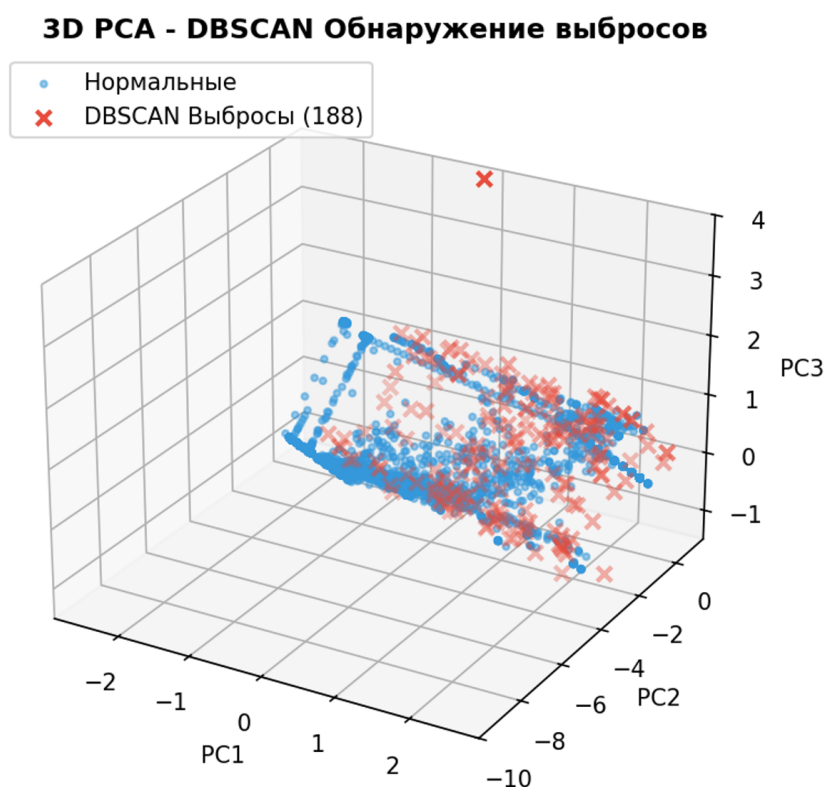


Рис. 6. Трёхмерная визуализация аномалий, обнаруженных методом DBSCAN

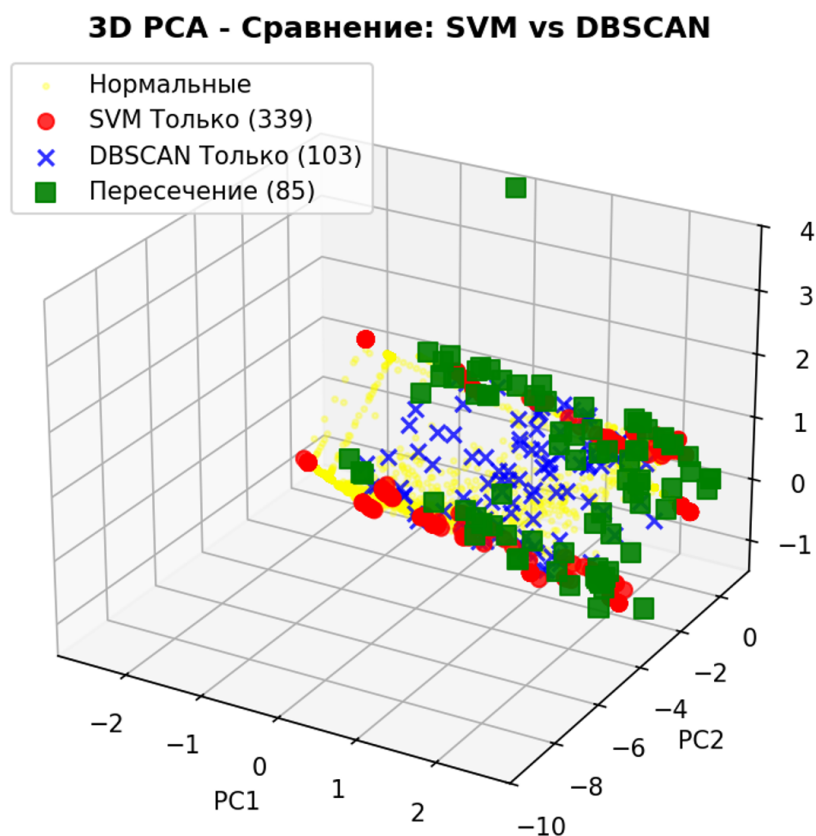


Рис. 7. Сравнение результатов обнаружения выбросов методами SVM и DBSCAN в трёхмерном пространстве главных компонент (PC1-PC2-PC3)

Дальнейшие исследования должны включать проверку подхода на других месторождениях, автоматическую настройку параметров и сопоставление выявленных выбросов с реальными инцидентами и данными технической диагностики.

Благодарности и финансирование

Исследование было проведено при поддержке РФФИ (проект № 24-11-00285).

The study was conducted with the support of the Russian Science Foundation (project No. 24-11-00285).

Список литературы

- [1] Ebrahimi, M. Anomaly detection in oil-producing wells: a comparative study of one-class classifiers in a multivariate time series dataset / M. Ebrahimi, E. Khamenechi, M. Ameri // *Journal of Petroleum Exploration and Production Technology*. – 2023. – Vol. 13. – P. 1845–1860.
- [2] Mohammadpour, A. Outlier detection and selection of representative fluid samples using machine learning: a case study of Iranian oil fields / A. Mohammadpour, D. Shahsavani, M. Riazi // *Journal of Petroleum Exploration and Production Technology*. – 2024. – Vol. 14, No. 10. – P. 2833–2852.
- [3] Liu, Y. Identification Method of Tubing Leakage Based on PCA-SVM in Gas Well / Y. Liu, H. Wang, Z. Chen // *Proceedings of the 2nd International Conference on Signal Processing and Intelligent Computing (SPIC)*. – 2024. – P. 1–6.
- [4] Wang, L. Identification method for true and false anomalies of oil pipeline leakage in an old industrial oil production well site / L. Wang, X. Zhang, J. Liu // *Oil & Gas Storage and Transportation*. – 2023. – Vol. 42, No. 5. – P. 567–575.
- [5] Comparison of Isolation Forest and One Class SVM in Anomaly Detection of Gas Pipeline Operation // *Proceedings of the 2023 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS)*. – IEEE, 2024. – (IEEE Xplore, 13 February 2024).
- [6] Reconstruction of well logging data using unsupervised machine learning based outlier detection techniques (UML ODTs) under adverse drilling conditions / J.-R. Chen, R.-Z. Yang, T.-T. Li [et al.] // *Applied Geophysics*. – 2025. – DOI 10.1007/s11770-025-1178-z.
- [7] Lundberg, S. M. A Unified Approach to Interpreting Model Predictions / S. M. Lundberg, S.-I. Lee // *Advances in Neural Information Processing Systems*. – 2017. – Vol. 30. – P. 4765–4774.
- [8] Ahmadi, M. Data-Driven Modeling of Oil Saturation from Dielectric Logs Using Ensemble Regression, Dimensionality Reduction and Anomaly Detection Machine Learning Algorithms / M. Ahmadi, M. Bagheri, A. Sadeghi // *SPE Annual Technical Conference and Exhibition*. – 2023. – Paper SPE-215078-MS.
- [9] Patel, D. Oil and gas flow anomaly detection on offshore naturally flowing wells using deep neural networks / D. Patel, M. Shah, K. Patel // *Journal of Petroleum Science and Engineering*. – 2024. – Vol. 234. – Art. 109876.
- [10] Hassan, N. Unsupervised machine learning model for predicting anomalies in subsurface safety valves and application in offshore wells during oil production / N. Hassan, A. El-Sayed, S. Ibrahim // *Journal of Petroleum Exploration and Production Technology*. – 2023. – Vol. 13. – P. 2211–2225.
- [11] Zhao, Q. Varying-scale HCA-DBSCAN-based anomaly detection method for multi-dimensional energy data in steel industry / Q. Zhao, M. Li, Y. Sun // *Information Sciences*. – 2023. – Vol. 645. – Art. 119479.

COMPARING SVM AND DBSCAN FOR OUTLIER DETECTION IN OIL WELL MONITORING DATA*

Mikhailov I.S.^{1,a}, Myo Hlaing Win^{1,b}, Sidorov K.O.^{1,c}

¹ National Research University "Moscow Power Engineering Institute" (NRU MPEI), Moscow, Russia.

E-mail: ^a fr82@mail.ru, ^b myohlaingwin69287@gmail.com, ^c kirill.sidoroff2014@yandex.ru

Abstract. The aim of this study is to compare One-Class Support Vector Machine (One-Class SVM) and DBSCAN for detecting statistical anomalies in oil-well monitoring data and to assess their joint use. Six parameters from a multiphase water-cut meter were analyzed; Principal Component Analysis (PCA) reduced them to three principal components preserving 89.86% of the total variance. In a sample of 8639 observations, One-Class SVM identified 424 anomalous points and DBSCAN identified 188, with 85 observations detected by both methods. For flow-regime classification by gas-content level, SVM with an RBF kernel ($C = 100.0$, $\gamma = 1.0$) achieved 98.26% accuracy with an F1 score of 0.97–0.99 on a stratified 70/30 split. Five-fold stratified cross-validation yielded an F1_macro of 0.9853 ± 0.0026 . Joint interpretation separates points detected by both methods from SVM-only and DBSCAN-only detections for subsequent expert review.

Keywords: outlier detection, Support Vector Machine, DBSCAN, Principal Component Analysis, oil well monitoring, machine learning, flow regime classification, anomalies.

References

- [1] Ebrahimi, M., Khamsehchi, E., & Ameri, M. (2023). Anomaly detection in oil-producing wells: A comparative study of one-class classifiers in a multivariate time series dataset. *Journal of Petroleum Exploration and Production Technology*, 13, 1845–1860.
- [2] Mohammadpour, A., Shahsavani, D., & Riazi, M. (2024). Outlier detection and selection of representative fluid samples using machine learning: A case study of Iranian oil fields. *Journal of Petroleum Exploration and Production Technology*, 14(10), 2833–2852.
- [3] Liu, Y., Wang, H., & Chen, Z. (2024). Identification method of tubing leakage based on PCA-SVM in gas well. In *Proceedings of the 2nd International Conference on Signal Processing and Intelligent Computing (SPIC)* (pp. 1–6).
- [4] Wang, L., Zhang, X., & Liu, J. (2023). Identification method for true and false anomalies of oil pipeline leakage in an old industrial oil production well site. *Oil & Gas Storage and Transportation*, 42(5), 567–575.
- [5] Comparison of isolation forest and one class SVM in anomaly detection of gas pipeline operation. (2024). In *Proceedings of the 2023 International Conference on Informatics, Multimedia, Cyber and Information System (ICIMCIS)*. IEEE.

For citation:

Mikhailov I.S., Myo Hlaing Win, Sidorov K.O. Comparing SVM and DBSCAN for Outlier Detection in Oil Well Monitoring Data. *Intelligent transport*, 2026, no. 3(39), pp. 53–68. DOI: 10.24412/3033-6007-2026-339-53-68. (in Russian)

*The study was conducted with the support of the Russian Science Foundation (project No. 24-11-00285).

- [6] Chen, J.-R., Yang, R.-Z., Li, T.-T., Xu, Y.-D., & Sun, Z.-P. (2025). Reconstruction of well logging data using unsupervised machine learning based outlier detection techniques (UML ODTs) under adverse drilling conditions. *Applied Geophysics*. <https://doi.org/10.1007/s11770-025-1178-z>
- [7] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 4765–4774.
- [8] Ahmadi, M., Bagheri, M., & Sadeghi, A. (2023). Data-driven modeling of oil saturation from dielectric logs using ensemble regression, dimensionality reduction and anomaly detection machine learning algorithms. In *SPE Annual Technical Conference and Exhibition* (Paper SPE-215078-MS).
- [9] Patel, D., Shah, M., & Patel, K. (2024). Oil and gas flow anomaly detection on offshore naturally flowing wells using deep neural networks. *Journal of Petroleum Science and Engineering*, 234, Article 109876.
- [10] Hassan, N., El-Sayed, A., & Ibrahim, S. (2023). Unsupervised machine learning model for predicting anomalies in subsurface safety valves and application in offshore wells during oil production. *Journal of Petroleum Exploration and Production Technology*, 13, 2211–2225.
- [11] Zhao, Q., Li, M., & Sun, Y. (2023). Varying-scale HCA-DBSCAN-based anomaly detection method for multi-dimensional energy data in steel industry. *Information Sciences*, 645, Article 119479.