**ГЕОФИЗИКА**  
 **Анализ данных экспериментов по термически стимулированному разрушению образцов песчаника с использованием машинного обучения**

*Индаков Г.С., студент MSU.AI*

*Беляков Н.В., преподаватель MSU.AI*

*Казначеев П.А., научный руководитель НКР*

**Аннотация**

В работе исследуется возможность выделения групп схожих импульсов термически стимулированной акустической эмиссии на основе параметров импульсов в эксперименте по нагреву образца метаморфизированного песчаника. Данный образец горной породы был подвергнут предварительному механическому нагружению, в результате чего в нем образовались крупные трещины. Предполагается, что импульсы, генерируемые вследствие процессов развития таких трещин, могут иметь схожие характеристики, благодаря чему их можно различить на общем фоне событий. Импульсы были предварительно сгруппированы в 3 класса на основании вычисления взаимнокорреляционной функции. Анализ полученных данных методами машинного обучения показывает, что без учета волновых форм импульсов различные классы оказываются слабо различимы на общем фоне событий, однако 3-й класс стабильно оказывается различим лучше других. Это может свидетельствовать о том, что импульсы 3-го класса действительно соответствуют группе схожих событий, в то время как другие два класса имеют помеховое происхождение.

**Ключевые слова**

горные породы, акустическая эмиссия, термическое воздействие, классификация, машинное обучение

**Введение**

Акустическая эмиссия (АЭ) в лабораторных экспериментах с образцами горных пород является одним из ключевых источников информации о протекании процессов разрушения в твердом теле. Импульсы АЭ являются упругими волнами, распространяющимися в материале при структурных изменениях от источников АЭ, таких как, например, образование и развитие трещин [1]. Каждый импульс характеризуется набором параметров, на основе которых можно судить о свойствах трещины и процессе ее развития. Совокупность данных об АЭ позволяет производить анализ статистики импульсов и выделять паттерны схожих событий.

Предполагается, что волновые формы импульсов одного и того же источника или схожих источников также схожи между собой. Мера их сходства обычно определяется корреляционной функцией, инвариантной к абсолютному значению амплитуды импульса и к сдвигу волновой формы во времени. Коэффициент корреляции часто определяется как максимум взаимнокорреляционной функции. Вычислительная сложность такого выделения групп импульсов АЭ составляет весьма большую величину порядка o(n2∙m2), где n – длина кадра регистрации (импульса), а m – число зарегистрированных импульсов. Для уменьшения затрат времени и вычислительных ресурсов на определение устойчивых групп импульсов АЭ представляется актуальным применение методов машинного обучения.

Целью работы является установление возможности выявления устойчивых групп импульсов и характерных особенностей процесса разрушения образцов горных пород на основе параметров волновых форм импульсов термостимулированной АЭ (ТАЭ). Статистические параметры импульсов ТАЭ могут коррелировать между собой, что имеет физическое обоснование для интегральных параметров разрушения, т.е. для величин, характеризующих отдельные эксперименты по нагреву с разными образцами [2]. Объектом изучения являются данные термического нагружения образцов горных пород, в первую очередь, волновых форм импульсов ТАЭ, такие как амплитуда, продолжительность сигнала, количество пересечений порога регистрации и др. Особенностью данного исследования является предварительное разбиение совокупности импульсов ТАЭ на группы импульсов на основании сходства их волновых форм, что не предоставляет исключительной гарантии принадлежности выбранных импульсов соответствующим классам. В связи с этим исследование включает в себя задачу кластеризации импульсов ТАЭ для определения их принадлежности различным группам, независимо от существующей разметки по классам.

В задачах анализа данных АЭ исследователей часто интересует мониторинг состояния материалов и конструкций, что включает в себя выделение полезного сигнала на фоне шумов, поиск источника дефектов и классификацию сигналов. Это можно сделать на основе спектральных характеристик акустического сигнала, который несет информацию о масштабах развивающихся дефектов [3]. В теоретическом отношении вопросы локации важны для сейсмологии в условиях отсутствия информации о скоростях упругих волн в среде [4]. Актуальной задачей для горнодобывающей отрасли является предсказание горных ударов на основе сигналов-предвестников АЭ. В работе [5] отмечено, что на основе таких параметров волновых форм импульсов АЭ, как амплитуда, длительность, энергия, средняя и пиковая частоты, удается выделять предвестниковые сигналы от трещин, инициирующих горный удар. Этой проблеме посвящены работы [6-7]. Авторы [7] используют нейронную сеть с 2 скрытыми слоями – полносвязным и регуляризации – с cross-entropy в качестве loss function. Достигнутые значения accuracy для распознавания сигналов АЭ от источников определенных типов, варьируются от 40 до 98%. Тем не менее, accuracy не всегда выступает лучшей метрикой. В работе [8] анализируются матрицы сигналов, составленные из волновых форм импульсов при пиковых значениях механической нагрузки, с целью классификации «полезных» импульсов на фоне шума. Для такого анализа авторы применяют CNN с оценкой таких метрик, как accuracy, recall, precision, f1-score, достигая показателей более 90%.

В целом изучению механического нагружения горных пород посвящено множество работ, в то время как влияние термического воздействия остается менее изученным. Тем не менее, в основе термической инициации трещин лежат механизмы, связанные с отсутствием преимущественной ориентации поля глобальных напряжений, создаваемой при механической нагрузке [9]. Таким образом, данное исследования направлено на анализ данных другой геофизической природы и представляет собой попытку выявления особенностей разрушения, связанных, главным образом, с тепловым воздействием. С теоретической точки зрения, это является актуальной задачей для расширения представлений о подготовке вулканических землетрясений, связанных с разогревом пород в вулканической камере. В прикладном отношении информация о поведении твердых материалов при воздействии высоких температур важна для строительства пожаростойких зданий и повышения надежности проектирования захоронений радиоактивных отходов, так как радиация является одним из главных источников разогрева пород при залегании [1].

**Данные**  
  
Были использованы данные эксперимента по нагреву предварительно одноосно нагруженного образца песчаника. Данные представляют собой таблицу параметров зарегистрированных импульсов ТАЭ. Фрагмент датасета представлен в таблице 1. Имеется 15 признаков и 1 целевая переменная – *Clust* или индекс группы. В ходе предварительной обработки с помощью взаимнокорреляционного метода было выявлено 3 группы импульсов (1, 2, 3 соответственно), остальные импульсы отнесены к классу 0. Таким образом, к классу 0 принадлежат те импульсы, которые имеют уникальные волновые формы и поэтому могут быть отнесены к группе фонового шума, состоящему из импульсов от многочисленных микротрещин и частично из помеховых импульсов.

Таблица 1. Фрагмент датасета

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Time** | **Aadc** | **AdB** | **AuV** | **A2hcounts** | **-//-** | **Clust** |
| 230,456 | 57 | 36,8 | 69,4 | 6,9 | … | 0 |
| 645,535 | 83 | 40,1 | 101,1 | 6,3 | … | 0 |
| 648,554 | 58 | 37 | 70,6 | 14,1 | … | 0 |
| 923,214 | 60 | 37,3 | 73,1 | 8,1 | … | 0 |
| 1608,312 | 34 | 32,3 | 41,4 | 41,4 | … | 0 |
| 1821,328 | 52 | 36 | 63,3 | 12,7 | … | 0 |
| … | … | … | … | … | … | … |

Из исходного датасета (3138 строк данных) были исключены импульсы, относящиеся к типичным помехам в соответствии с условиями, налагаемыми на среднюю частоту *AF* и амплитудно-временной параметр сигнала *RA*: 50 < AF < 550 кГц, 1 < RA < 100 В/с. Также был удален признак *Aadc*, представляющий собой амплитуду импульса в единицах АЦП и линейно связанный с амплитудой в мкВ – признаком *AuV*.

В фильтрованном датасете представлено 14 признаков и 2053 строк данных, причем подавляющее большинство импульсов принадлежит фоновому классу. Полная характеристика классов по числу импульсов представлена на рисунке 1-а. Таким образом, в распределении событий по классам датасет является существенно несбалансированным. Некоторые признаки показывают высокую корреляцию друг с другом, что отражено на рисунке 1-б. Тем не менее, из-за недостатка данных и слабой чувствительности моделей деревьев к скоррелированным признакам оставим большинство признаков. Корреляция между некоторыми из них может отражать характеристику волновой формы, так как по физической смыслу коррелирующие признаки могут быть принципиально различны, как, например, *Tdur* и *hcounts*.

|  |  |
| --- | --- |
|  |  |
| а) | б) |
| Рис.1 – Фильтрованный датасет: а) – количество импульсов в классах, б) – тепловая карта взаимной корреляции признаков. | |

**Методология**  
  
Предварительно был проведен дисперсионный анализ признаков датасета для проверки различимости классов. Была исследована совокупность признаков с помощью многомерного дисперсионного анализа (MANOVA) [10], затем оценено распределение классов по каждому признаку с помощью одномерного дисперсионного анализа (ANOVA) [11]. Для проверки различий классов между собой было проведено их попарное сравнение при помощи теста Тьюки [12-13].

Для анализа датасета с целью классификации импульсов были использованы стандартные методы ML: AdaBoost [14], Random forest [15], XGBoost [16], CatBoost [17]. Данные делились на обучающую и тестовую выборку со стратификацией в соотношении 65% к 35% ввиду малого числа импульсов для тестирования модели. К данным применялись разные способы уравнивания числа объектов разных классов для обучения: undersampling, oversampling, SMOTE [18] и собственным методом на основе гауссового шума в качестве погрешности параметра (1% величины). Лучший результат получен с применением SMOTE. Наиболее подходящие параметры для модели подбирались с помощью Optuna [19]. Визуализация распределения данных по классам в пространстве признаков с помощью LDA для исходных данных и для тренировочной части датасета приведена на рисунке 2. Проводилась кросс-валидация (от 5 до 30 выборок), оценивались метрики f1-score, precision, recall.

|  |  |
| --- | --- |
|  |  |
| а) | б) |
| Рис.2 – Распределения данных по классам в пространстве признаков (LDA): а) – исходный датасет, б) – тренировочная часть датасета, аугментированная SMOTE. | |

Так как классы 1, 2 и 3 малочисленны и представляют собой «полезный» сигнал на общем фоне импульсов, был проведен поиск аномалий в данных на полном датасете (unsupervised) с использованием моделей Isolation forest [20], One-class SVM [21] и LOF [22] c подбором параметра contamination/nu с помощью Optuna. В качестве метрики использовалось число импульсов классов 1-3, распознанных алгоритмом как аномалия. Кластеризация выделенных аномалий производилась методами K-means [23] и DBSCAN [24] с целью получения исходного разбиения импульсов на классы 1-3.

**Дисперсионный анализ признаков**

С помощью MANOVA было проанализировано влияние номера класса на вектор из 13 зависимых переменных, что меньше на 1 признак от общего числа содержащихся в датасете из-за высокой корреляции между некоторыми признаками. Анализировались четыре тестовых критерия: Wilks’ lambda, Pillai’s trace, Hotelling–Lawley trace и Roy’s greatest root [10]. Каждый из них оценивает гипотезу о равенстве векторных средних во всех группах, определяемых номером класса. Результаты, представленные в таблице 2, показывают, что классы статистически различимы: номер класса импульса оказывает статистически значимое влияние на многомерный профиль признаков. Тем не менее, доля объяснённой дисперсии в критериях, используемых MANOVA, составляет менее 10%, что указывает на умеренную силу различий между классами.

Таблица 2. Результаты MANOVA

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Clust** | **Value** | **Num DF** | **Den DF** | **F Value** | **Pr > F** |
| Wilks' lambda | 0.9061 | 13.0000 | 2039.0000 | 16.2510 | 0.0000 |
| Pillai's trace | 0.0939 | 13.0000 | 2039.0000 | 16.2510 | 0.0000 |
| Hotelling-Lawley trace | 0.1036 | 13.0000 | 2039.0000 | 16.2510 | 0.0000 |
| Roy's greatest root | 0.1036 | 13.0000 | 2039.0000 | 16.2510 | 0.0000 |

ANOVA подтверждает выводы MANOVA (таблица 3): из 14 проанализированных признаков, 10 оказались статистически значимыми при уровне значимости 0.05 [11]. Несмотря на то, что различия между классами 1-3 могут быть выражены неярко, эти классы существенно отличимы от фонового класса 0, что особенно заметно для класса 3. На рисунке 3 приведены распределения некоторых признаков по классам.

Таблица 3. Результаты ANOVA: значимые признаки (p < 0.05)

|  |  |  |  |
| --- | --- | --- | --- |
| **№** | **Feature** | **F-statistic** | **p-value** |
| 1 | AdB | 28.998149 | 2.33E-18 |
| 2 | EdB | 22.181315 | 3.92E-14 |
| 3 | AuV | 8.808881 | 8.40E-06 |
| 4 | hcounts | 7.865258 | 3.23E-05 |
| 5 | Emarse | 7.677510 | 4.22E-05 |
| 6 | Trise | 5.860930 | 5.54E-04 |
| 7 | A2hcounts | 5.727658 | 6.69E-04 |
| 8 | AF | 3.554993 | 1.38E-02 |
| 9 | Tdur | 3.064379 | 2.71E-02 |
| 10 | Heat | 2.655802 | 4.70E-02 |

|  |  |
| --- | --- |
|  |  |
| а) | б) |
| Рис.3 – Распределения некоторых признаков по классам: а) – *AdB* (амплитуда в дБ), б) – *hcounts* (количество пересечений порога). | |

Результаты парных сравнений признаков по тесту Тьюки [13] показывают, что распределение признаков между классами неоднородно и имеет четко выраженную иерархию отличий. Для большинства признаков есть уверенное различие между классами 0 и 2, 0 и 3, 1 и 2, 1 и 3. При этом между классами 0 и 1 по большинству признаков значимых различий не наблюдается (*AdB*, *EdB*, *AuV*, *hcounts*, *Emarse*, *Trise*), а класс 3, как правило, отличается от всех остальных классов по большинству признаков с высокой статистической значимостью. Визуализация теста Тьюки для некоторых признаков приведена на рисунке 4.

|  |  |
| --- | --- |
|  |  |
| а) | б) |
| Рис.4 – Тест Тьюки для некоторых параметров: а) – *AdB* (амплитуда в дБ), б) – *hcounts* (количество пересечений порога). | |

Таким образом, классы 0 и 1 могут рассматриваться как близкие или смежные группы, а классы 2 и 3 - как более отчетливо разделенные. Следует ожидать, что при классификации классы 2 и 3 будут легче определяться, причем класс 3 лучше прочих.

**Результаты**  
  
В данном разделе приведены основные показатели моделей, описанных в методологии.

1. AdaBoost

Адаптивный бустинг использовал глубину дерева max\_depth < 3, так как ее увеличение приводило к сильному замедлению процесса обучения без улучшения метрик. Классы 2 и 3 распознавались моделью, класс 1 не был распознан. Классификационный отчет модели приведен в таблице 4. Матрица несоответствий для модели AdaBoost, подобранной Optuna на кросс-валидации с числом фолдов 5 и оценкой recall, и вклад признаков показаны на рисунке 5.

Таблица 4. AdaBoost: Classification report

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **class** | **precision** | **recall** | **F1-score** | **support** |
| 0 | 0.99 | 0.98 | 0.98 | 706 |
| 1 | 0.00 | 0.00 | 0.000 | 2 |
| 2 | 0.12 | 0.14 | 0.13 | 7 |
| 3 | 0.33 | 0.50 | 0.40 | 4 |

|  |  |
| --- | --- |
|  |  |
| а) | б) |
| Рис.5 – AdaBoost model: а) – матрица несоответствий, б) – вклад признаков. | |

2. Random forest

Основным параметром, влияющим на модель, являлась максимальная глубина дерева. Однако при любой глубине класс 3 выделялся в целом лучше: 2/4 для max\_depth выше 5, в то время как класс 1 не был выделен при любой конфигурации, а оценки для класса 2 оказывались неустойчивыми. На рисунке 6-а показана матрица несоответствий для модели Random forest, подобранной Optuna на кросс-валидации с числом фолдов 5 и оценкой recall, в таблице 5 приведен классификационный отчет модели. На рисунке 6-б отображен вклад признаков.

Таблица 5. Random forest: Classification report

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **class** | **precision** | **recall** | **F1-score** | **support** |
| 0 | 0.987 | 0.987 | 0.987 | 706 |
| 1 | 0.000 | 0.000 | 0.000 | 2 |
| 2 | 0.250 | 0.286 | 0.267 | 7 |
| 3 | 0.400 | 0.500 | 0.444 | 4 |

|  |  |
| --- | --- |
|  |  |
| а) | б) |
| Рис.6 – Random forest model: а) – матрица несоответствий, б) – вклад признаков. | |

3. XGBoost

Модель XGBoost лучше всего определяла класс 3 из трех целевых классов при любой конфигурации параметров. В таблице 6 и на рисунке 7 приведены результаты тестирования модели с параметрами, подобранными с помощью Optuna на кросс-валидации с числом фолдов 30 и оценкой recall.

Таблица 6. XGBoost: Classification report

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **class** | **precision** | **recall** | **F1-score** | **support** |
| 0 | 0.99 | 0.99 | 0.99 | 706 |
| 1 | 0.00 | 0.00 | 0.00 | 2 |
| 2 | 0.38 | 0.43 | 0.40 | 7 |
| 3 | 0.43 | 0.75 | 0.55 | 4 |

|  |  |
| --- | --- |
|  |  |
| а) | б) |
| Рис.7 – XGBoost model: а) – матрица несоответствий, б) – вклад признаков. | |

4. CatBoost

Модель существенным образом не отличается от предыдущих, показывая те же скромные результаты. В таблице 7 и на рисунке 8 показаны результаты тестирования модели CatBoost, подобранной с помощью Optuna с оценкой recall для валидационной выборки (20% от обучающей выборки) вместо кросс-валидации для экономии времени на обучение. В отличие от прочих моделей, CatBoost набольшее внимание обращает на бинарный признак *Heat* (рисунок 8-б), определяющий этап нагрева, что может свидетельствовать о значимости учета температурных данных.

Таблица 7. CatBoost: Classification report

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **class** | **precision** | **recall** | **F1-score** | **support** |
| 0 | 0.99 | 0.97 | 0.98 | 706 |
| 1 | 0.00 | 0.00 | 0.00 | 2 |
| 2 | 0.23 | 0.43 | 0.30 | 7 |
| 3 | 0.29 | 0.50 | 0.36 | 4 |

|  |  |
| --- | --- |
|  |  |
| а) | б) |
| Рис.8 – CatBoost model: а) – матрица несоответствий, б) – вклад признаков. | |

5. Isolation forest

Модель использовалась в двух вариантах: с делением выборки на обучающую/тестовую и без разделения.

На основе обучающей выборки с помощью Optuna были подобраны оптимальные параметры модели на основании метрики f1-score, так как при выделении аномалий существенным является также то, что модель может выделить вместе с импульсами целевых классов 1-3 большое число фоновых импульсов. Optuna установила параметр contamination, отвечающий за процент зашумления (аномалий) в данных равным 0.1884. В таблице 8 приведены результаты тестирования модели, выраженные в количестве импульсов классов, распознанных как аномальные. Процент выделения класса 0 в качестве аномалии можно считать погрешностью работы алгоритма.

Таблица 8. Isolation forest: результаты на тестовой выборке

|  |  |  |  |
| --- | --- | --- | --- |
| **Класс** | **Выделено импульсов** | **Всего импульсов в классе** | **% выделенных импульсов от общего числа импульсов в классе** |
| 0 | 129 | 706 | 18 % |
| 1 | 0 | 2 | 0 % |
| 2 | 3 | 7 | 43 % |
| 3 | 3 | 4 | 75 % |

Во втором случае Isolation forest применялся ко всему датасету. Визуализация выделенных аномалий приведена на рисунке 9. Видно, что алгоритм старается выбрать значения, находящиеся на границе облака точек по некоторым признакам, в то время как изначальная классификация проводилась с учетом волновых форм и может быть не отражена в признаковом пространстве, полученном для анализа.

В таблице 9 приведен результат работы Isolation forest с параметром contamination = 18%. В то время как для классов 1 и 2 алгоритм справляется слабо, класс 3 выделяется практически целиком.

|  |
| --- |
|  |
| Рис.9 – Isolation forest: аномалии (весь датасет, аномальные данные отмечены синими крестиками. |

Таблица 9. Isolation forest: результаты на всем датасете

|  |  |  |  |
| --- | --- | --- | --- |
| **Класс** | **Выделено импульсов** | **Всего импульсов в классе** | **% выделенных импульсов от общего числа импульсов в классе** |
| 0 | 350 | 2015 | 17 % |
| 1 | 0 | 7 | 0 % |
| 2 | 10 | 21 | 48 % |
| 3 | 9 | 10 | 90 % |

6. One-Class SVM

Аналогично Isolation forest был применен SVM. Для сравнения работы разных методов выявления аномалий, был установлен одинаковый параметр зашумления (18%). Результаты приведены в таблице 10. Алгоритм хуже распознает 3 класс, хотя этот класс все равно выделяется лучше остальных.

Таблица 10. One-Class SVM: результаты на всем датасете

|  |  |  |  |
| --- | --- | --- | --- |
| Класс | Выделено импульсов | Всего импульсов в классе | % выделенных импульсов от общего числа импульсов в классе |
| 0 | 357 | 2015 | 18 % |
| 1 | 0 | 7 | 0 % |
| 2 | 6 | 21 | 28 % |
| 3 | 5 | 10 | 50 % |

7. Local Outlier Factor (LOF)

В отличие от двух предыдущих методов, LOF обращает большее внимание на первый класс. Однако его общие результаты сравнительно хуже, чем у SVM и Isolation forest. Результаты работы LOF приведены в таблице 11.

Таблица 11. LOF: результаты на всем датасете

|  |  |  |  |
| --- | --- | --- | --- |
| Класс | Выделено импульсов | Всего импульсов в классе | % выделенных импульсов от общего числа импульсов в классе |
| 0 | 362 | 2015 | 18 % |
| 1 | 3 | 7 | 43 % |
| 2 | 4 | 21 | 19 % |
| 3 | 1 | 10 | 10 % |

8. Кластеризация: K-means и DBSCAN

К аномалиям, выделенным с помощью Isolation forest, как лучшего среди испробованных методов, были применены стандартные методы кластеризации – K-means и DBSCAN. Так как классы 1 и 2 были представлены несколькими импульсами, проводилось разбиение аномалий на два класса – фон и «полезный» сигнал, в который, как предполагалось, могут попасть импульсы классов 1-3. Визуализация разбиения классов в пространстве признаков, редуцированном к двумерному представлению с помощью PCA, приведена на рисунке 10. К сожалению, в выделенный минорный кластер не попали импульсы из классов 1-3, а попали 5 импульсов фонового класса.

|  |  |
| --- | --- |
|  |  |
| а) | б) |
| Рис.10 – Кластеризация аномалий на 2 кластера: а) – K-means, б) – DBSCAN. | |

**Выводы**

В ходе выполнения работы были применены различные методы машинного обучения для задач классификации и поиска аномалий в данных с последующей кластеризацией. Сравнение результатов показывает, что несмотря на относительно слабую классификацию/кластеризацию импульсов на основе разметки по классам, полученным в ходе предварительного взаимнокорреляционного анализа, на основе параметров импульсов ТАЭ удается наблюдать статистически значимое различие классов между собой. Результаты дисперсионного анализа позволяют заключить, что классы 0 и 1 не различимы по подавляющей части признаков, а импульсы классов 2 и 3 устойчиво отличаются от фоновых событий, причем класс 3 отличается наиболее значимо от всех прочих классов. Алгоритмы машинного обучения выделяют импульсы, принадлежащие классу 3, лучше остальных. Класс 2 выделяется хуже, а класс 1 практически не выделяется. При этом многие фоновые импульсы (класс 0) классифицируются алгоритмами как аномальные/целевые. Это может быть связано с дисбалансом классов и техническими аспектами применения машинного обучения, но также это может свидетельствовать в пользу того, что импульсы класса 3, скорее всего, действительно связаны между собой, что отражается в параметрах ТАЭ, в то время как импульсы классов 1 и 2 не имеют устойчивой связи и могли быть выделены взаимнокорреляционным методом ошибочно. Кроме того, некоторые фоновые импульсы могут быть связаны друг с другом и с импульсами классов 1-3, о чем свидетельствует тенденция различных алгоритмов выделять их как аномалии. Таким образом, применение машинного обучения для анализа данных ТАЭ, представленных параметрами импульсов ТАЭ, позволяет на качественном уровне валидировать данные взаимнокорреляционного анализа волновых форм сигналов.

**Код и данные для воспроизведения результатов**

Ссылка на репозиторий GitHub:

<https://github.com/Gutleyb/MSU.AI-Sandstone_ML/tree/main>

**Список литературы**

[1] Шкуратник В.Л., Новиков Е.А., Вознесенский А.С., Винников В.А. Термостимулированная акустическая эмиссия в геоматериалах. — M. : Горная книга, 2015. — 241 с. — ISBN: 978-5-98672-401-0.

[2] Индаков Г.С., Казначеев П.А. Оценка статистических параметров потока импульсов термически стимулированной акустической эмиссии в лабораторных экспериментах // Ученые записки физического факультета Московского университета. 2021. № 1:2110501.

[3] Ciaburro G., Iannace G. Machine-Learning-Based Methods for Acoustic Emission Testing: A Review. Appl. Sci. 2022, 12, 10476. https://doi.org/10.3390/app122010476

[4] Zhao Q., Glaser S.D. Relocating Acoustic Emission in Rocks with Unknown Velocity Structure with Machine Learning. Rock Mech. Rock. Eng. 2020, 53, 2053–2061. https://doi.org/10.1007/s00603-019-02028-8

[5] Sun J., Liu D., He P., Guo L., Cao B., Zhang L., Li Z., Experimental investigation on acoustic emission precursor of rockburst based on unsupervised machine learning method. Rock Mechanics Bulletin, 2024, v. 3-2, 100099, ISSN 2773-2304, https://doi.org/10.1016/j.rockmb.2023.100099

[6] Liu X., Liang Z., Zhang Y., Wu X., Liao Z. Acoustic Emission Signal Recognition of Different Rocks Using Wavelet Transform and Artificial Neural Network // *Shock and Vibration*. 2015, 846308, 14 p. https://doi.org/10.1155/2015/846308

[7] Константинов А.В., Рассказов И.Ю. Разработка комплекса нейросетевых моделей для идентификации типа источника акустического излучения на удароопасном месторождении // Горный информационно-аналитический бюллетень. – 2024. – № 11. – С. 23–36. DOI: 10.25018/0236\_1493\_2024\_11\_0\_23

[8] Xie P., Chen W., Zhao W., Gao H. Acquisition of acoustic emission precursor information for rock masses with a single joint based on clustering-convolutional neural network method // Journal of Rock Mechanics and Geotechnical Engineering. 2024, v.16-12. 5061-5076 Pp., ISSN 1674-7755, https://doi.org/10.1016/j.jrmge.2024.01.016

[9] Соболев Г.А. Модель лавинно-неустойчивого трещинообразования – ЛНТ // Физика Земли. — 2019. — № 1. — С. 166–179.

[10] Johnson R.A., Wichern D.W. Applied Multivariate Statistical Analysis. 6th Edition, Pearson Prentice Hall, Upper Saddle River. 2007

[11] Field A. Discovering Statistics Using IBM SPSS Statistics: And Sex and Drugs and Rock “N” Roll, 4th Edition, Sage, Los Angeles, London, New Delhi. 2013

[12] James G., Witten D., Hastie T., Tibshirani R. An Introduction to Statistical Learning / with Applications in R. 2nd Edition. Springer New York, NY. 2021. https://doi.org/10.1007/978-1-0716-1418-1

[13] Tukey J.W. A Survey of Sampling from Contaminated Distributions. In: Oklin, I., Ed., Contributions to Probability and Statistics, Stanford University Press, Redwood City, CA. 1960

[14] Freund Y., Schapire R.E. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting // Journal of Computer and System Sciences. – 1997 – Vol.55, Issue 1 – Pp. 119-139. https://doi.org/10.1006/jcss.1997.1504.

[15] Breiman L. Random Forests: Machine Learning // Kluwer Academic Publishers. Netherlands – 2001 – Vol. 45, Issue 1, Pp. 5- 32 https://doi.org/10.1023/A:1010933404324

[16] Chen T., Guestrin C. XGBoost: A Scalable Tree Boosting System // In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16). Association for Computing Machinery, New York, NY, USA. – 2016 – Pp. 785–794. https://doi.org/10.1145/2939672.2939785

[17] Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. V., & Gulin, A. CatBoost: unbiased boosting with categorical features. In Advances in Neural Information Processing Systems. / Curran Associates, Inc. – 2018 – Vol. 31.

https://proceedings.neurips.cc/paper\_files/paper/2018/file/14491b756b3a51daac41c24863285549-Paper.pdf

[18] Chawla N.V., Bowyer K.W., Hall L.O., Kegelmeyer W.P. SMOTE: Synthetic Minority Over-sampling Technique // Journal of Artificial Intelligence Research. – 2002 – Vol.16 – Pp. 321-357. https://doi.org/10.1613/jair.953

[19] Akiba T. et al. Optuna: A next-generation hyperparameter optimization framework // Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining. – 2019. – С. 2623-2631.

[20] Liu F.T., Ting K.M., Zhou Z.-H., "Isolation Forest," 2008 Eighth IEEE International Conference on Data Mining, Pisa, Italy, 2008, pp. 413-422, doi: 10.1109/ICDM.2008.17.

[21] Manevitz L.M., Yousef M. One-class SVMs for document classification //Journal of machine Learning research. – 2001. – Vol. 2. – Dec. – Pp. 139-154.

[22] Breunig M.M., Kriegel H.P., Ng R.T., Sander J. LOF: identifying density-based local outliers. In Proceedings of the 2000 ACM SIGMOD international conference on Management of data. – 2000, May – Pp. 93-104.

[23] MacQueen J. Some methods for classification and analysis of multivariate observations //Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics. – University of California press, 1967. – Vol. 5. – Pp. 281-298.

[24] Ester M., Kriegel H. P., Sander J., Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise // In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD'96). AAAI Press. – 1996 – Pp. 226–231.