

Применение методов машинного обучения для классификации кварцитов по химическому составу: влияние микроэлементов и геохимическая идентификация

А. С. Мясникова, Р. Ю. Шендрик, И. А. Елисеев, О. И. Чачанагова,
А. М. Федоров, А. И. Непомнящих

Институт геохимии им. А.П. Виноградова СО РАН, 664033, г. Иркутск, ул. Фаворского, 1А,
e-mail: sasham@igc.irk.ru

Поступила в редакцию 30.12.2024 г., принята к печати 04.04.2025 г.

Объект исследования. Кварциты из различных участков Восточно-Саянского кварценосного района. **Цель.** Целью данного исследования является применение методов машинного обучения для эффективной классификации образцов кварцитов по их химическому составу, включая идентификацию ключевых микроэлементов, таких как марганец, и выявление геохимических различий между образцами. **Материалы и методы.** В исследовании использовались данные химического анализа 776 образцов кварцита, которые были подвергнуты интерпретации с помощью методов машинного обучения. В качестве методов были применены стандартные техники предварительной обработки данных, такие как нормализация, а также аугментация данных с использованием SMOTE для решения проблемы дисбаланса классов. В итоге был выбран алгоритм CatBoost, который показал высокую точность классификации. **Результаты.** Результаты кросс-валидации показали, что алгоритм CatBoost достиг точности классификации до 97%. Важность признаков указывает на то, что марганец является ключевым элементом в классификации образцов, в то время как такие элементы, как алюминий и калий, оказывают вспомогательное влияние. Также успешно проведен анализ классификации по цвету кварцитов с точностью до 0.94. **Выводы.** Исследование демонстрирует эффективность применения методов машинного обучения для анализа химического состава кварцитов, предоставляя новые возможности для геохимических и археологических исследований.

Ключевые слова: кварциты, машинное обучение, классификация горных пород, RobustScaler, статистический анализ

Источник финансирования

Исследование проведено в рамках выполнения государственного задания по Проекту № 0284-2021-0004

Application of machine learning methods to classify quartzites by chemical composition: the influence of trace elements and geochemical identification

Alexandra S. Myasnikova, Roman Yu. Shendrik, Igor A. Eliseev, Olga I. Chachanagova,
Alexander M. Fedorov, Alexander I. Nepomniyschikh

A.P. Vinogradov Institute of Geochemistry, SB RAS, 1A Favorsky st., Irkutsk 664033, Russia, e-mail: sasham@igc.irk.ru

Received 30.12.2024, accepted 04.04.2025

Research subject. Quartzites from Different Sites of the East Sayan Quartz-Bearing Region. **Aim.** The aim of this study is to apply machine learning methods to effectively classify quartzite samples by their chemical composition, including the identification of key trace elements such as manganese and the detection of geochemical differences between samples. **Materials and methods.** The study used chemical analysis data from 776 quartzite samples, which were interpreted using machine learning methods. Standard data preprocessing techniques such as normalization were applied as well as data augmentation using SMOTE to solve the class imbalance problem. As a result, the CatBoost algorithm was selected, which showed high classification accuracy. **Results.** Cross-validation results showed that the CatBoost algorithm

Для цитирования: Мясникова А.С., Шендрик Р.Ю., Елисеев И.А., Чачанагова О.И., Федоров А.М., Непомнящих А.И. (2025) Применение методов машинного обучения для классификации кварцитов по химическому составу: влияние микроэлементов и геохимическая идентификация. *Литосфера*, **25**(2), 320–335. <https://doi.org/10.24930/2500-302X-2025-25-2-320-335>. EDN: XFIXKQ

For citation: Myasnikova A.S., Shendrik R.Yu., Eliseev I.A., Chachanagova O.I., Fedorov A.M., Nepomniyschikh A.I. (2025) Application of machine learning methods to classify quartzites by chemical composition: the influence of trace elements and geochemical identification. *Lithosphere (Russia)*, **25**(2), 320–335. (In Russ.) <https://doi.org/10.24930/2500-302X-2025-25-2-320-335>. EDN: XFIXKQ

© А.С. Мясникова, Р.Ю. Шендрик, И.А. Елисеев, О.И. Чачанагова, А.М. Федоров, А.И. Непомнящих, 2025

achieved classification accuracy of up to 97%. The importance of features indicates that manganese is a key element in the classification of samples, while elements such as aluminum and potassium have a supporting effect. The analysis of the classification by color of quartzites with an accuracy of 0.94 was also successfully carried out. *Conclusions.* The study demonstrates the effectiveness of applying machine learning methods to the analysis of the chemical composition of quartzites, providing new opportunities for geochemical and archaeological research.

Keywords: *quartzites, machine learning, rock classification, RobustScaler, statistical analysis*

Funding information

The study was performed by the governmental assignment in terms of Project 0284-2021-0004

ВВЕДЕНИЕ

В последние годы методы машинного обучения становятся все более востребованными в геохимических исследованиях, предлагая новые возможности для анализа многомерных данных. Особый интерес представляет задача классификации горных пород по химическому составу, которая имеет значение как для фундаментальной геологии, так и для прикладных археологических исследований. В то время как традиционные статистические методы часто оказываются недостаточно эффективными для обработки сложных геохимических данных, алгоритмы машинного обучения демонстрируют впечатляющие результаты в подобных задачах, что подтверждается работами последних (Wang, 2021; Zhu, 2023; Hnila, 2025).

Особую актуальность приобретает изучение кварцитов – метаморфических пород, широко распространенных, но сложных для диагностики из-за вариабельности их химического состава. В геологическом аспекте точная классификация кварцитов позволяет уточнить условия их формирования и выявить закономерности распределения в пределах месторождений (Аюржанаева, 2020). Для археологии же определение источников каменного сырья имеет принципиальное значение при реконструкции древних производственных цепочек, торговых путей и миграций населения (Pitblado, 2008; Pitblado, 2013). Однако, в отличие от хорошо изученных в этом отношении обсидианов и кремней, кварциты до сих пор остаются недостаточно исследованными, особенно на территории России.

Данное исследование направлено на разработку методики дифференциации кварцитов Восточно-Саянского района с использованием современных методов машинного обучения. В работе решается комплекс взаимосвязанных задач: во-первых, оценивается возможность точной классификации образцов с разных участков одного месторождения на основе их геохимических характеристик; во-вторых, проводится сравнительный анализ эффективности различных алгоритмов для реше-

ния этой задачи; в-третьих, определяются ключевые химические маркеры, наиболее значимые для различения кварцитов. Особое внимание уделяется интерпретируемости моделей с помощью современных методов explainable AI (Lundberg, 2017), что важно как для геологов, так и для археологов.

Полученные результаты имеют междисциплинарное значение. С геологической точки зрения, они позволяют уточнить особенности химического состава кварцитов изучаемого региона и разработать новые подходы к их классификации. В археологическом аспекте исследование создает основу для будущих работ по атрибуции каменных артефактов и реконструкции древних хозяйственных систем Южной Сибири. Методика, предложенная в работе, может быть адаптирована для изучения других типов горных пород и применена при решении широкого круга задач в области геохимии и археометрии.

В исследовании наряду с определением месторождения была проведена классификация кварцитов по цвету для оценки возможностей машинного обучения в решении подобных задач и выявления связи между химическим составом и окраской пород. Этот аспект работы имеет важное практическое значение, поскольку цвет является одним из ключевых диагностических признаков, используемых как в археологической практике при атрибуции артефактов, так и в промышленности при сортировке минерального сырья. Однако автоматизация цветовой классификации сталкивается с рядом методологических сложностей, включая субъективность визуального описания, нелинейный характер зависимости между содержанием элементов-хромофоров и наблюдаемым цветом, а также влияние текстурных особенностей породы на ее оптические свойства.

ИСХОДНЫЕ ДАННЫЕ

Проявления кварцевого сырья Восточно-Саянского кварценосного района пространственно приурочены к карбонатно-кремнистой толще среднерифейского возраста в обрамлении Гарганской

глыбы. Мощность толщи достигает 1000 м, протяженность – 100–120 км с перерывами, от верховьев р. Ока до р. Онот. Кварцитовая толща имеет мощность не более 300 м. Карбонатно-кремнистая толща прорывается крупными телами гранитоидов (преимущественно гранодиоритами) сумсунурского комплекса возрастом 790 млн лет.

Кварциты толщи относятся к метаморфизованным хемогенно-осадочным образованиям. Кварциты представляют собой в различной степени перекристаллизованные массивные и пятнисто-полосчатые черные, темно-серые до белых, мелко- и тонкозернистые кварциты, темно-серые и серые сливные кварциты, белые, желтоватые, светло-серые и серые порфиробластические суперкварциты и их аналоги.

Пробы для исследований отбирались на Ока-Урикском и Урда-Гарганском участках (Федоров, 2021) расположенных на расстоянии 15 км друг от друга и разделенных Гарганским плутоном сумсунурского комплекса. Ока-Урицкий и Урда-Гарганский узлы представляют собой группы несколько изолированных друг от друга выходов на поверхность кварцитов, разделенных межгорными долинами. Ока-Урицкий и Урдагарганский узлы характеризуются однотипным набором разновидностей кварцитов, но с несколько отличными соотношениями их объемов.

Отличаются Ока-Урицкий и Урдагарганский узлы структурными особенностями строения толщи и, объемами и составом прорывающих толщу мелких интрузивных тел. Этим участкам свойственны открытые асимметричные складки, плоскость складчатости которых полого погружается в северо-западном направлении. Возникновение складчатости связано с межслойными надвигами и обусловлено продольным сжатием по линии северо-запад – юго-восток и наличием близко расположенного фундамента Гарганской глыбы.

Для Ока-Урикского узла характерны редкие и мелкие (до 3 м в поперечнике) дайки северо-западного, субширотного и субмеридионального простирания и штоки, не превышающие первых десятков метров в поперечнике. Интрузивные тела представлены гранодиоритами, плагиогранитами, кварцевыми диоритовыми порфиритами и тонкозернистым габбро.

В пределах Урдагарганского узла картируется крупная складка, рассчитанный наклон шарнира которой имеет падение в юго-западном направлении под углом около 17°. В апикальной части складки отмечается серия крупных интрузий риолитов, риодацитов, дацитов, граносиенитов, диоритов, габбро-диоритов, габбро с порфирами. Кроме того, отмечаются единичные маломощные пегматитовые жилы. Морфология интрузивных тел в плане варьирует от отдельных линз разной мощности до ветвящихся удлиненных даек. Протя-

женность интрузивных тел от первых десятков до 200 м, мощность от первых метров до 30 м. Простирание тел преимущественно северо-восточное, в редких случаях северо-западное.

Контактовое воздействие интрузий выражается в окварцевании, мусковитизации, калишпатизации, тремолитизации, метасоматизме и скарнировании карбонатных, кремнисто-карбонатных и кремневых пород на глубину до 10 м.

Пробы кварцитов анализировались в Институте геохимии им. А.П. Виноградова СО РАН методами масс-спектрометрии с индуктивно связанной плазмой (ICP-MS), пламенной фотометрии (ПФ) и атомного эмиссионно-спектрального анализа (АЭСА). Методика предварительной пробоподготовки подробно описана в работах (Fedorov, 2019; Непомнящих, 2023), поэтому в данной статье мы на этом останавливаться не будем. В результате исследований было проанализировано 254 пробы с Ока-Урикского участка и 522 пробы с Урда-Гарганского узла. Данные были объединены и перемешаны для целей изучения возможности определения месторождения по данным химического анализа (далее – сводная таблица). Кроме того, был проведен анализ данных по отдельности для разных участков. Примесный состав исследуемых кварцитов приведен на рис. 1. Как видно из этого рисунка, примесный состав этих участков в среднем отличается, однако отнести случайно найденный образец кварцита к какому-либо участку затруднительно в связи с большим разбросом примесного состава.

В литературе есть данные об определении месторождений и генезиса кварцитов на основе их химического состава (Götze, 2012; Müller, 2012; Pitblado, 2013; Shah, 2022), однако, в образцах, взятых с Ока-Урикского и Урда-Гарганского узлов не наблюдается четкого распределения примесей, что хорошо видно на рис. 2.

На рис. 3 представлены корреляционные зависимости примесного состава кварцитов Ока-Урикского и Урда-Гарганского участков, рассчитанные с помощью коэффициента Φ_K . Φ_K – это коэффициент корреляции, который последовательно применяется к категориальным, порядковым и интервальным переменным. Он основан на уточнении теста гипотезы Пирсона о независимости переменных и интерпретируется как наклон в повернутом двумерном нормальном распределении. Преимущества Φ_K включают согласованное применение к различным типам переменных, учет нелинейной зависимости, а также восстановление коэффициента корреляции Пирсона для двумерного нормального распределения. Это делает данный коэффициент полезным для анализа корреляционных матриц с переменными смешанного типа. Подробности о методологии и оценке статистической значимости корреляций можно найти

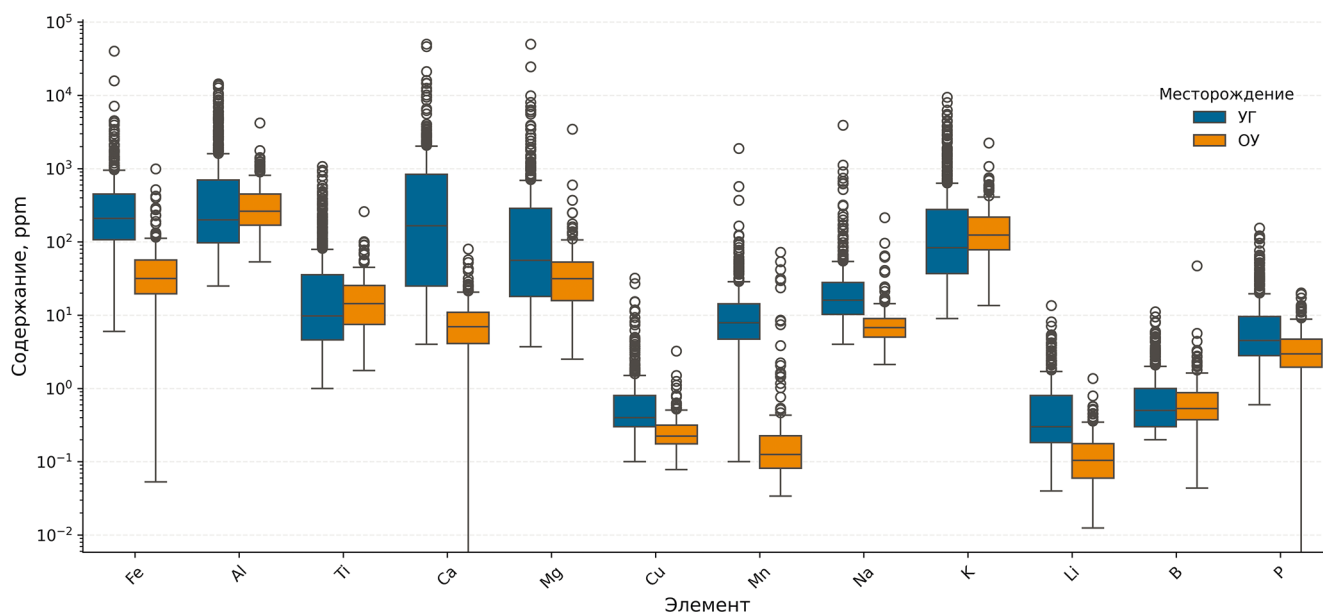


Рис. 1. Карта содержания элементов в кварцитах Ока-Урикского узла (ОУ) и Урда-Гарганского узла (УГ).

Высота цветных столбиков отражает диапазон квартилей. Черные границы представляют собой средние значения, в то время как горизонтальная черная линия внутри каждого цветного столбика обозначает медиану. “Усы” каждой боксовой диаграммы указывают на экстремальные значения, которые находятся в пределах 1.5 раз межквартильного размаха за пределами краев столбика. Белые точки соответствуют “выбросам”.

Fig. 1. Elemental content map of quartzites from the Oka-Urik cluster (OY) and Urda-Gargan cluster (UG).

The height of the colored bars reflects the quartile range. The black borders represent the mean values, while the horizontal black line inside each colored bar denotes the median. The whiskers of each box plot indicate extreme values that are within 1.5 times the interquartile range outside the bar edges. White dots correspond to outliers.

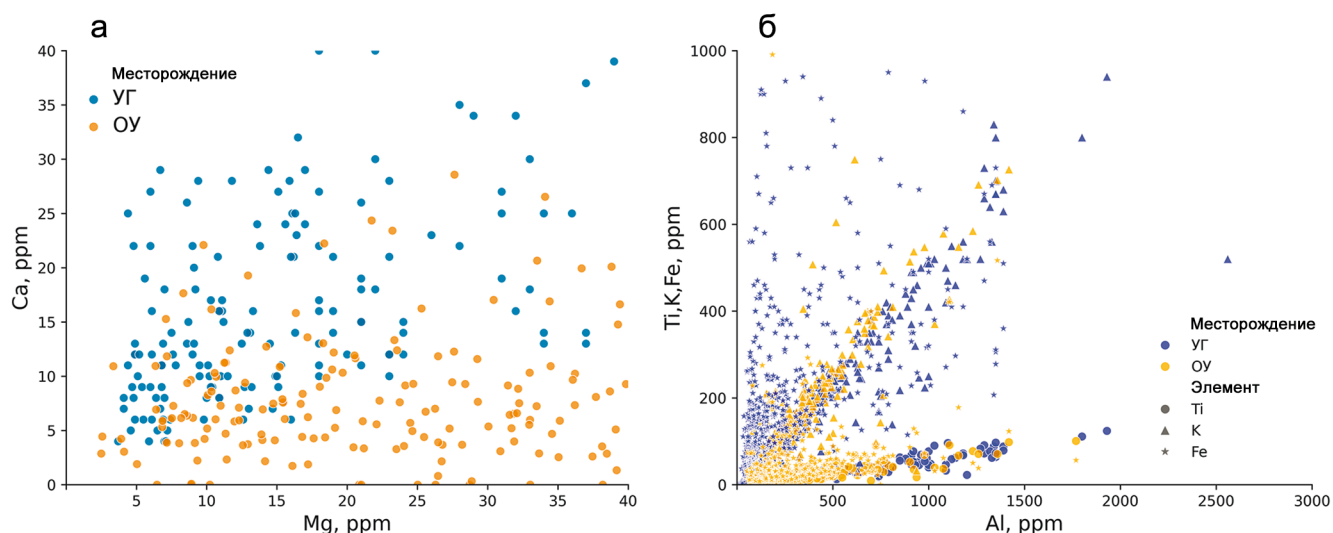


Рис. 2. Двумерные графики концентраций элементов в образцах кварцитов: концентрации кальция и магния (а); сводные концентрации титана, калия и железа в зависимости от концентрации алюминия (б).

Данные приведены для всех образцов кварцитов. УГ – кварциты Урда-Гарганского участка, ОУ – кварциты Ока-Урикского участка.

Fig. 2. Two-dimensional plots of elemental concentrations in quartzite samples: calcium and magnesium concentrations (a); combined titanium, potassium, and iron concentrations versus aluminum concentrations (b).

Data are presented for all quartzite samples. УГ – quartzites of the Urda-Gargan area, ОУ – quartzites of the Oka-Urik area.

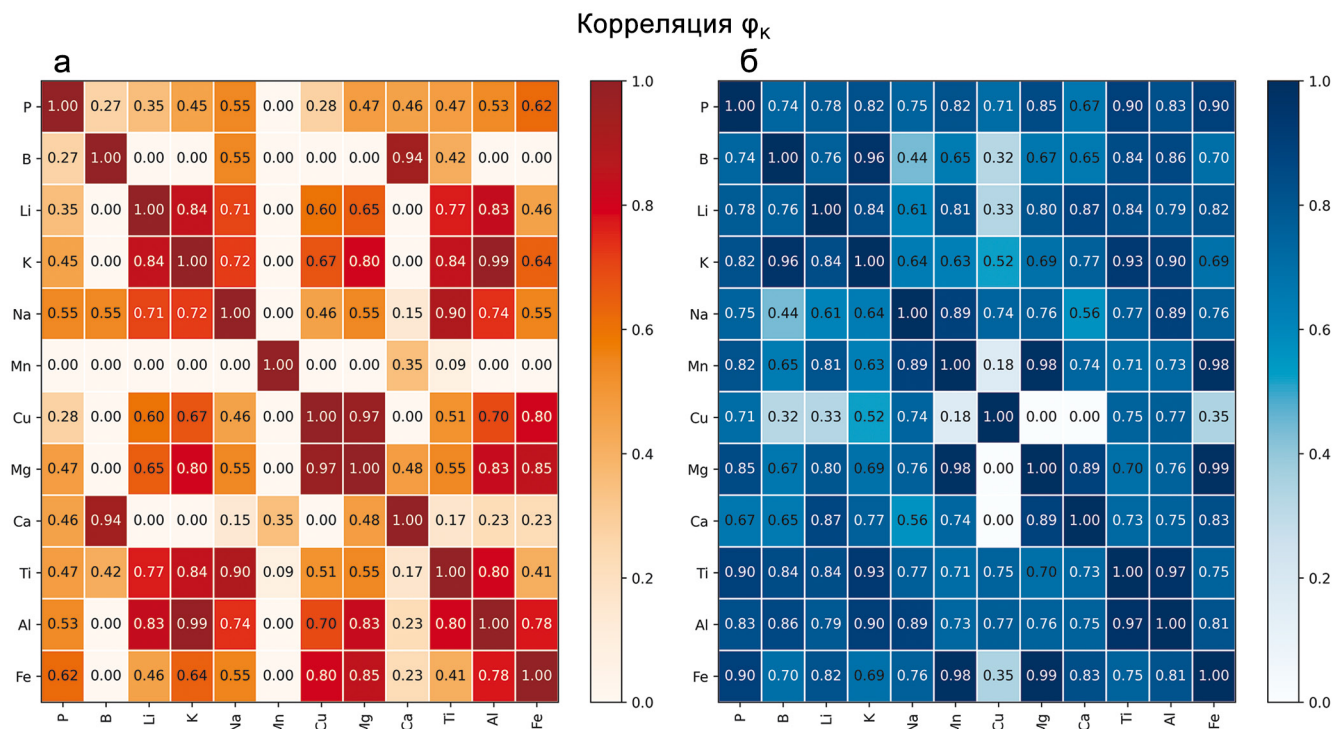


Рис. 3. Карты корреляции по Φ_K между элементами для кварцитов Ока-Урикского узла (а) и Урда-Гарганского узла (б).

Чем темнее цвет – тем сильнее корреляционная зависимость.

Fig. 3. Φ_K correlation maps between elements for quartzites of the Oka-Urik cluster (a) and the Urda-Gargan cluster (б).

The darker the color, the stronger the correlation.

в соответствующей публикации (Ваак, 2020). Оценка карт позволяет сделать некоторые выводы о взаимном распределении примесей:

– для Ока-Урикского узла характерны сильные корреляционные зависимости между Al и K, а также между Ca и B;

– для Урда-Гарганского узла наблюдается сильная корреляция между такими элементами как Al и Ti, Al и K, Fe, Mn и Mg.

Необходимо отметить, что матрицы корреляции (даже учитывающие нелинейные зависимости) не дают полной картины о взаимосвязи элементов в кварцитах, так как учитывают только их попарное распределение. Далее мы покажем, что методы машинного обучения могут улавливать более тонкие зависимости распределения элементов.

В ходе исследования цветовая классификация кварцитов проводилась визуальным методом без использования стандартизированных цветовых шкал. Все образцы были разделены на цветовые группы на основе непосредственного визуального осмотра, что соответствует практике многих полевых и лабораторных исследований в геологии и археологии. Такой подход, несмотря на опреде-

ленную субъективность, оправдан в рамках данного исследования по нескольким причинам.

1. Основной целью цветовой классификации была проверка принципиальной возможности связи макроскопических визуальных характеристик с химическим составом, а не создание эталонной системы цветовых стандартов.

2. Визуальная оценка остается наиболее распространенным и практически значимым методом в археологической практике при работе с каменными артефактами.

3. Все образцы оценивались одним исследователем в сходных условиях освещения, что обеспечивало относительную однородность оценок.

Однако у данного метода остались ограничения, связанная с некоторой условностью границ между цветовыми группами. Также не учитывались возможные оптические эффекты, связанные с текстурой образцов. Кроме того, тонкие цветовые нюансы могли остаться незамеченными.

Тем не менее, даже такая упрощенная классификация позволила выявить статистически значимые связи между визуальными характеристиками и химическим составом, что подтверждает практи-

ческую ценность проведенного анализа. Полученные результаты следует рассматривать как первый шаг в изучении этой проблемы, открывающий перспективы для более точных измерений с использованием спектрофотометрических методов в будущих исследованиях.

МЕТОДЫ ИССЛЕДОВАНИЯ

Для статистического анализа геохимических различий между образцами кварцитов из разных участков были последовательно применены два метода. Тест Манна-Уитни, как непараметрический аналог t -критерия, позволил выявить значимые различия в распределениях концентраций элементов между Ока-Урикским и Урда-Гарганским участками, несмотря на наличие выбросов и отклонения от нормального распределения в данных. Критерий Дипера дополнил этот анализ, обнаружив внутреннюю неоднородность в распределениях некоторых элементов, что проявлялось в мультимодальности их концентрационных профилей. Алюминий показал статистически значимые различия между участками ($p = 0.042$), что может быть связано с различным содержанием слюд в кварцитах. Бор и титан демонстрируют тенденцию к различиям ($p = 0.090$ и $p = 0.087$ соответственно), однако эти результаты не достигают стандартного уровня значимости. Остальные элементы проявляют исключительно сильные различия ($p < 0.001$), подтверждая их роль как надежных геохимических маркеров. Полученные статистические выводы стали основой для последующего применения методов машинного обучения, подтвердив как значимые различия между участками, так и сложную внутреннюю структуру данных, требующую учета при построении классификационных моделей.

Первоначально была исследована возможность разделения образцов по химическому составу с помощью простых классических методов. К сожалению, метод главных компонент (англ. *principal component analysis*, PCA) в данном случае оказался неспособен отобразить принципиальные различия в химическом составе образцов с двух разных участков. Однако, похожий с PCA метод t -SNE достаточно хорошо справился с данной задачей (рис. 4), что дало шанс на то, что более мощные методы машинного обучения также справятся с определением месторождения по химическому составу. t -SNE (*t-distributed Stochastic Neighbor Embedding*) – это алгоритм машинного обучения для нелинейного снижения размерности, разработанный для визуализации высокоразмерных данных в двух- или трехмерном пространстве, моделируя похожие объекты как близкие точки, а непохожие – как отдаленные (Maaten, 2014). Метод t -SNE способен обнаружить нелинейные зависимости, что эффективно для выявления сложных паттернов, когда ли-

нейные методы не дают результатов. Кроме того, этот метод сохраняет локальную структуру данных, что удобно для анализа кластеров и визуализации объектов, находящихся вблизи друг от друга в исходном пространстве. Этот метод особенно полезен для нахождения скрытых паттернов в данных.

Сам процесс машинного обучения состоял из нескольких этапов, так как предварительная обработка данных играет ключевую роль в оптимизации их качества. Процессы, такие как нормализация, преобразование и расширение набора данных, значительно повышают точность и эффективность последующей классификации. Это особенно важно в контексте машинного обучения, поскольку качество входных данных оказывает прямое влияние на продуктивность модели. Во-первых, была выполнена стандартизация и нормализация данных. В работе для нормализации геохимических данных был применен метод *RobustScaler*, обеспечивающий устойчивость к выбросам и значительным вариациям в концентрациях химических элементов. В отличие от стандартных методов нормализации, использующих среднее значение и стандартное отклонение, *RobustScaler* основан на медиане и межквартильном размахе (IQR), что делает его особенно подходящим для анализа данных с аномальными значениями и нестандартными распределениями. Преобразование данных выполняется по формуле, где из каждого значения вычитается медиана соответствующего признака, после чего результат делится на межквартильный размах (разность между 75-м и 25-м перцентилями). Такой подход сохраняет структуру исходных данных и обеспечивает сопоставимость различных химических элементов, концентрации которых могут отличаться на несколько порядков. Выбор *RobustScaler* обусловлен спецификой геохимических данных, для которых характерны широкий диапазон концентраций (от следовых количеств до процентов), наличие выбросов и непараметрические распределения. Метод демонстрирует устойчивость к аномальным значениям, не искажая общую картину распределения данных, что критически важно для последующего применения алгоритмов машинного обучения, чувствительных к масштабу признаков. В частности, *RobustScaler* позволяет корректно сравнивать вклад различных элементов, концентрации которых изначально измерялись в разных единицах или имели существенно отличающиеся диапазоны значений. Практическая реализация нормализации включала следующие этапы: сначала для каждого химического элемента вычислялись медиана и межквартильный размах на основе обучающей выборки, затем эти параметры использовались для преобразования всех данных. Важно отметить, что параметры нормализации определялись только на обучающей выборке, а затем применялись к тестовым данным, что исключает инфор-

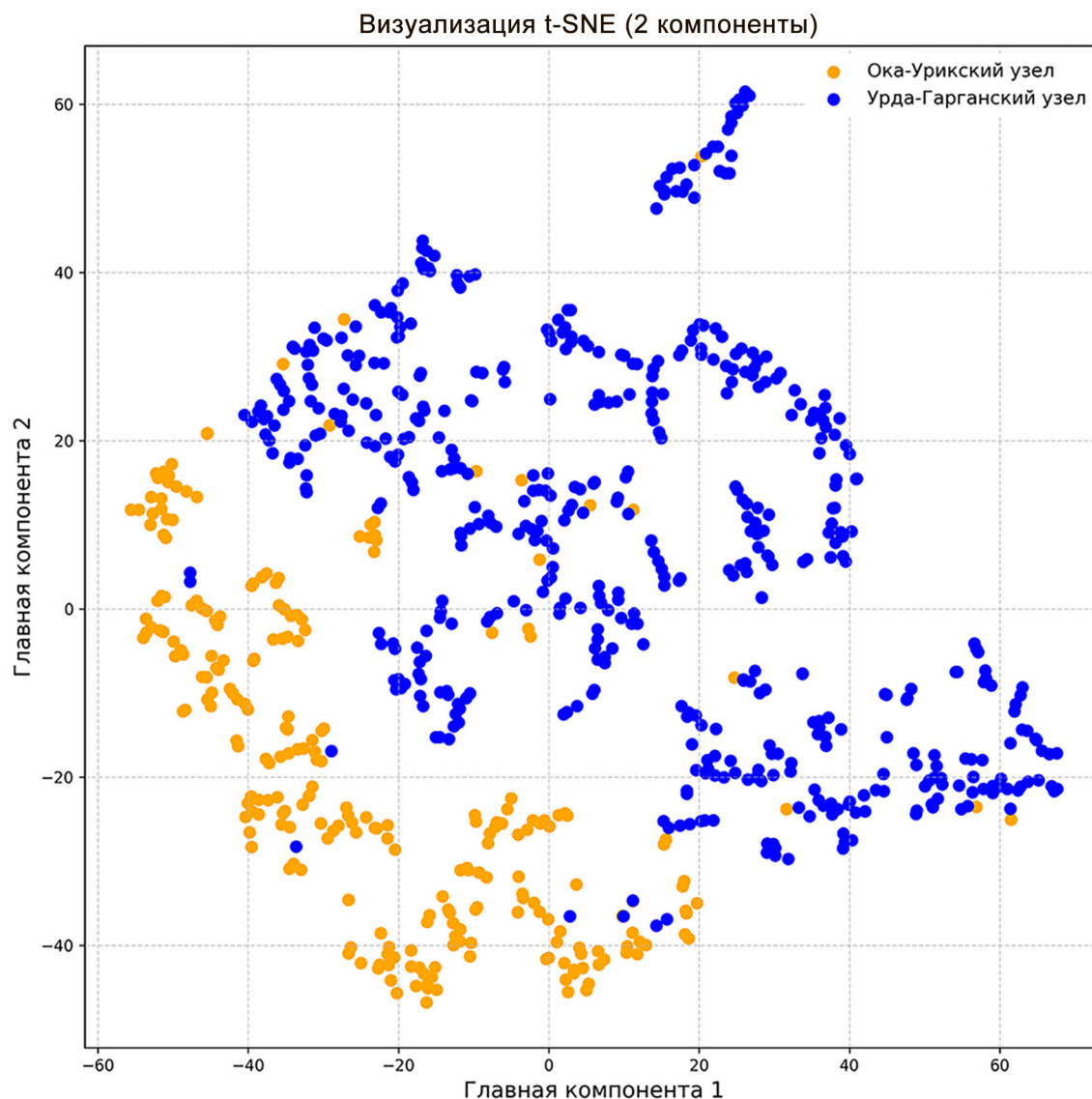


Рис 4. 2-х компонентное распределение t_SNE по сводной таблице.

Визуализация методом t-SNE выполнена с параметрами: размерность проекции – 2D, Perplexity – 10 (оптимизировано по метрике `silhouette_score`), random state – 12345 (для воспроизводимости).

Fig. 4. 2-component t_SNE distribution for the summary table.

Visualization using the t-SNE method was performed with the following parameters: projection dimension – 2D, Perplexity – 10 (optimized by the `silhouette_score` metric), random state – 12345 (for reproducibility).

мационную утечку и обеспечивает корректность валидации моделей. Визуальный анализ распределений до и после нормализации подтвердил эффективность метода: RobustScaler успешно уменьшил влияние выбросов, сохранив при этом основные особенности исходных данных. Полученные результаты свидетельствуют о целесообразности использования RobustScaler для предобработки геохимических данных перед применением методов машинного обучения. Это особенно полезно для алгоритмов машинного обучения, которые

чувствительны к масштабу, таким как градиентный спуск или алгоритмы на основе расстояний (например, K-ближайших соседей).

Следующим шагом стала аугментация данных для расширения проблемы дисбаланса данных. В качестве метода реализации аугментации данных был выбран алгоритм SMOTE (Chawla, 2002). На рис. 5 представлены результаты десятикратной перекрестной проверки для пяти методов и разных целевых переменных, причем данные иллюстрируют как результаты до, так и после примене-

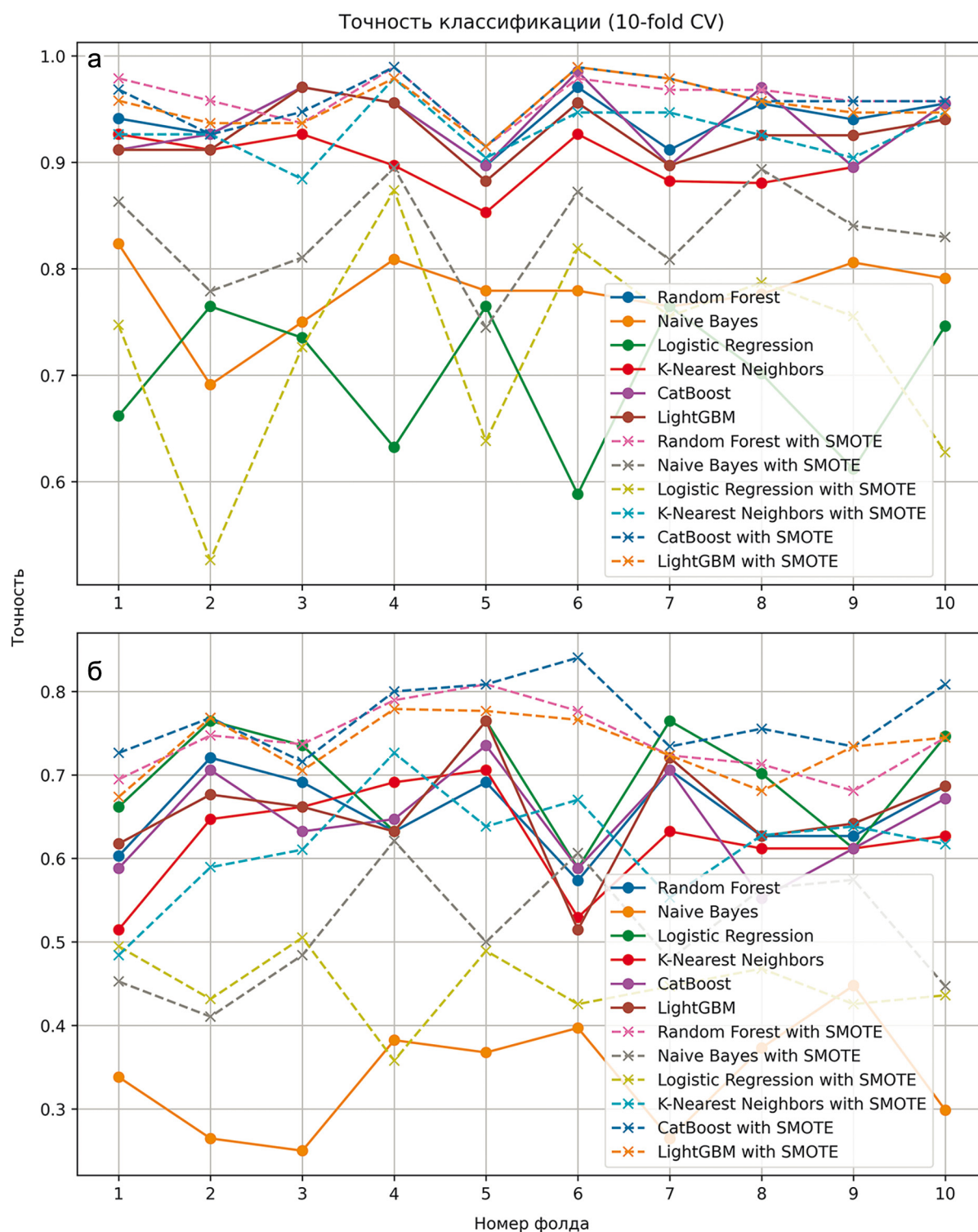


Рис. 5. Кривые кросс-валидации с n-кратным разделением для разных целевых переменных: месторождение (а) и цвет (б).

Fig. 5. Cross-validation curves with n-fold splitting for different target variables: deposit (a) and color (б).

ния SMOTE к набору данных. Использование десятикратной перекрестной проверки имеет широкое применение, так как позволяет снизить неопреде-

ленность в процессе выбора тестового набора данных. Однако применение SMOTE не привело к статистически значимому улучшению классификации

по месторождению (Дассурасу = +2.3%). Это может объясняться несколькими причинами: во-первых, исходный дисбаланс классов не требовал коррекции, во-вторых, решающее значение имел элемент, распределение которого уже было сбалансировано, и, в-третьих, синтетические образцы искажали корреляции между второстепенными элементами. Таким образом, для данного типа геохимических данных с умеренным дисбалансом предобработка SMOTE может быть избыточной. Однако, применение SMOTE при решении задачи классификации по цвету приводит к значительному увеличению работоспособности модели и увеличению точности классификации.

Здесь необходимо отметить, то методы, основанные на градиентном бустинге (CatBoost, LightGBM, XGBoost), демонстрируют высокую устойчивость к неоднородности данных, разбросу значений и дисбалансу классов. Тем не менее, в соответствии с общепринятой практикой, мы также включили в анализ базовую модель (например, Naive Bayes и Logistic Regression) в качестве контрольного алгоритма для валидации результатов.

Одним из важных моментов в машинном обучении является выбор классификатора. В нашем исследовании мы использовали следующие классификаторы.

1. *Random Forest* (Breiman, 2001). Ансамблевый метод, который использует множество деревьев решений для классификации. Каждое дерево обучается на случайной подвыборке данных, что помогает уменьшить риск переобучения и повышает общую точность модели.

2. *Naive Bayes* (Manning, 2008). Вероятностный классификатор, основанный на теореме Байеса, который предполагает независимость признаков. Он идеально подходит для задач, где разнообразие данных и классификация по категориям имеют важное значение, например, в спам-фильтрах.

3. *Logistic Regression* (Yu, 2011). Этот метод используется для бинарной классификации, прогнозируя вероятность принадлежности к классу. Несмотря на название, это на самом деле метод линейной модели, который применяет логистическую функцию для преобразования линейной комбинации входных признаков в вероятность.

4. *K-Nearest Neighbors (KNN)*. Метод, основанный на принципе близости, который классифицирует новые данные, основываясь на классах k ближайших соседей. Этот метод прост в реализации и эффективен для небольших наборов данных, но может быть медлительным для крупных выборок.

5. *CatBoost* (Prokhorenkova, 2017) и *LightGBM* (Ke, 2017). Градиентные бустинги, разработанные для работы с категориальными признаками без необходимости их предварительной обработки. Они идеально подходят для обработки смешанных типов данных и могут эффективно управлять боль-

шими объемами информации, обеспечивая высокую точность модели.

В качестве метрики оценки классификаторов была использована точность (ассигасу), которая представляет собой долю правильно классифицированных примеров к общему числу примеров в тестовом наборе. Эта метрика позволяет получить общее представление о производительности модели, но может быть недостаточно информативна в случае сильно несбалансированных данных. Однако, так как мы используем балансировку данных, то считаем эту метрику достаточной для наших целей.

Настройка алгоритмов машинного обучения проводилась с использованием фреймворка Optuna v3.0.0 (<https://optuna.org>), позволяющего эффективно подбирать оптимальные комбинации гиперпараметров. В ходе исследования проводился поиск по сетке с байесовской оптимизацией, что существенно сокращает время вычислений по сравнению с полным перебором. Для каждого классификатора был определен уникальный набор настраиваемых параметров и диапазоны их значений, учитывающие специфику алгоритмов. В случае Random Forest оптимизировались количество деревьев, глубина, минимальное число образцов для разделения и другие ключевые параметры. Для логистической регрессии подбирались коэффициент регуляризации, тип нормы и метод оптимизации. Градиентный бустинг требовал настройки скорости обучения, глубины деревьев и параметров регуляризации.

Проведенная оптимизация гиперпараметров выявила несколько характерных закономерностей. Для ансамблевых методов (Random Forest, CatBoost, LightGBM) наблюдается устойчивая тенденция к использованию относительно большого количества базовых estimators (300–500), что соответствует современным рекомендациям по построению композитных моделей. При этом глубина деревьев сознательно ограничивается (5–9 уровней), что свидетельствует о направленности на предотвращение переобучения через регуляризацию.

Особый интерес представляет согласованность параметров, контролирующих сложность деревьев:

- минимальное число образцов для разделения (`min_samples_split = 3` в Random Forest)
- минимальное число образцов в листе (`min_samples_leaf = 1`)
- ограничение максимального количества признаков (`max_features = 'log2'`)

Эти значения указывают на баланс между сохранением достаточной выразительной мощности моделей и контролем их склонности к переобучению. В алгоритмах, основанных на градиентном бустинге (CatBoost, LGBM, XGBoost) дополнительно прослеживается единообразие в выборе скорости обучения (`learning_rate 0.1–0.14`), что со-

ответствует общепринятому “золотому диапазону” для данного класса методов. Для линейных моделей (Logistic Regression) и метода k-ближайших соседей параметры подобраны в соответствии с их теоретическими ограничениями. Особенно показателен крайне малый коэффициент сглаживания ($\text{var_smoothing} = 2.21\text{e-}11$) в наивном байесовском классификаторе, что говорит о работе с данными высокой размерности. Общей чертой всех моделей стало использование умеренных значений регуляризационных параметров (reg_alpha , reg_lambda), что подтверждает сбалансированность обучающей выборки и отсутствие необходимости в жестких ограничениях.

Основываясь на данных кросс-валидации (см. рис. 5) для дальнейшей работы был выбран алгоритм CatBoost показавший хорошую точность в обоих случаях.

Анализ важности признаков, влияющих на целевую метрику, производился, как с помощью встроенных методов, так и с помощью библиотеки SHAP (Lundberg, 2017). Анализ важности признаков позволяет углубить понимание распределения данных и повысить интерпретируемость модели классификации. В этом исследовании для анализа важности признаков был использован метод SHAP, который основывается на теории игр и вычисляет влияние отдельных признаков на целевую переменную с учетом полного набора признаков. Среднее абсолютное значение влияния признаков может служить индикатором их значимости, при этом результаты не зависят от используемой модели.

РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

Как было указано ранее, методы машинного обучения на основе градиентного спуска (CatBoost и LightGBM) показали лучшую метрику на кросс-валидации, равную 0.97. Однако такую же метрику показала и древовидная модель случайного леса (Random Forest), что говорит о том, что данные участки достаточно хорошо разделяются по химическому составу, несмотря на их достаточно близкое расположение и мы можем с точностью до 97% отнести любой исследуемый образец к той или иной группе. Если посмотреть на диаграммы важности признаков, то определяющим элементом в наших исследованиях является марганец, причем его влияние на этот процесс достаточно велико (рис. 6). Здесь необходимо отметить, что данные у нас нормированы и методы машинного обучения исследуют не абсолютные значения величин примесей, а именно их взаимное соотношение. Следовательно, существует значительная разница в распределении Mn в кварците исследуемых двух участков. Интересно, что влияние содержания остальных элементов практически одинаково. На диаграмме SHAP можно выделить та-

кую особенность, что для В, Al и К показана обратная зависимость – их наибольшие значения указывают на принадлежность к Ока-Урикскому участку, тогда как высокие значения Mn, Li, Ca и P определяют принадлежность к Урда-Гарганскому узлу. Такое распределение скорее всего связано с минералами, которые окружают зерна кварца в кварцитах. Как было показано ранее для Ока-Урикского узла характерна высокая корреляция Al и К, что могло бы говорить о вхождении ионов алюминия в зерна кварца с ионами калия в качестве компенсаторов заряда. Однако проведенные ЭПР исследования не показывают никаких сигналов, свойственных ионам алюминия, поэтому можно говорить о том, что Al и К находятся в составе слюды, скорее всего мусковита. Для Урда-Гарганскому участка же характерна связь кальция и фосфора, что характерно для фосфоритов, а также корреляция Mn-Fe-Mg, что скорее всего говорит о наличии родонита. Отметим, что и фосфориты, и родонит могли образоваться в результате метаморфизма осадочных пород, как и кварциты исследуемых участков, однако никем не были обнаружены в кварцитах Восточного Саяна. Однако обнаружен апатит $\text{Ca}_5(\text{PO}_4)_3$ как связующее звено между кальцием и фосфором. Данный аспект требует более глубокого изучения.

Однако при внимательном рассмотрении диаграмм важности признаков на рис.6 можно увидеть разное взаимное расположение элементов по важности. Сравнительный анализ важности признаков с использованием SHAP-значений и метода `feature_importances_` в CatBoost требует понимания их принципиальных различий. SHAP-значения, основанные на теории кооперативных игр, оценивают вклад каждого признака с учетом всех возможных взаимодействий между ними, предоставляя информацию как о величине, так и о направлении влияния на прогноз. В отличие от этого, `feature_importances_` в CatBoost отражает частоту использования признаков при построении деревьев и степень улучшения качества разделения, не учитывая направленность воздействия. На практике согласованные результаты обоих методов для определенных признаков (например, содержания Mn) подтверждают их ключевую роль в модели. Расхождения же могут указывать на разные аспекты влияния: высокая SHAP-важность при низком показателе `feature_importances_` характерна для признаков, действующих через сложные взаимодействия, тогда как обратная ситуация свидетельствует о частом использовании признака для расщеплений при его малом непосредственном вкладе в итоговый прогноз.

На рис. 5 было показано, что методы машинного обучения с неплохой точностью (до 0.94) можно также использовать для исследования цвета кварцитов. Здесь необходимо отметить, что низкая точность методов может быть связана еще и описанием

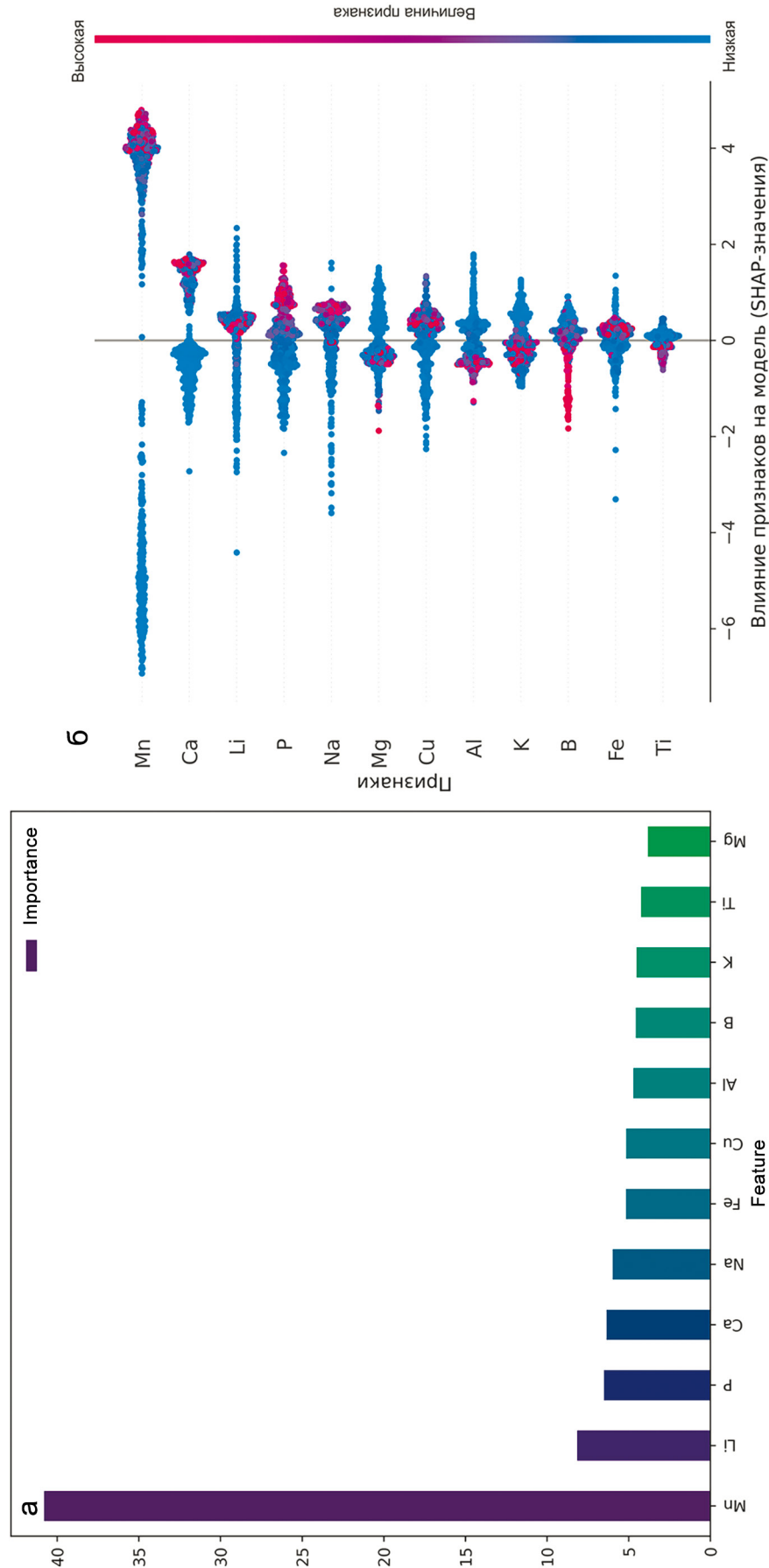


Рис. 6. Диаграммы важности признаков, рассчитанные с использованием метода feature_importances_ в CatBoost (а) и SHAP-значений (б).

Целевая переменная – месторождение. Признаки расположены в порядке убывания важности на основе их средних значений.

Fig. 6. Feature importance diagrams calculated using the feature_importances_ method in CatBoost (a) and SHAP values (б).

The target variable is deposit. Features are arranged in descending order of importance based on their average values.

цвета кварцитов (разметкой данных), так как это величина, которая зависит от многих факторов, таких как освещенность, время суток, осадки и т. д. Применение методов предобработки данных, таких как RobustScaler, хотя и обеспечило устойчивость к выбросам, могло непреднамеренно нивелировать слабые, но потенциально значимые геохимические сигналы, особенно для элементов с низкими концентрациями. Тем не менее на расширенном наборе данных были рассчитаны важности признаков с помощью метода SHAP. Результаты представлены на рис. 7. Из диаграмм видно, что на идентификацию цвета кварцитов примеси влияют по-разному. Так, идентификация белых кварцитов связана с большими значениями Li и Ca и низкими значениями Mn. Для светло-серого кварцита будет характерно повышенное содержание Na, K и Ti. Для черных и темно серых кварцитов определяющим является высокое содержание Mn и Fe, что в принципе согласуется с предположением, что темную окраску кварцитам дают соединения марганца.

Отметим, что корреляционные карты для кварцитов, разделенных по цвету, практически одинаковы и не дают возможности сделать какие-либо выводы.

Еще одним примером того, что анализ важности факторов в методах машинного обучения работает лучше корреляции, является анализ концентрации кальция. На рис. 8 показаны диаграммы важности признаков для задачи регрессии по определению концентрации кальция на основе концентрации других примесей для кварцитов с разных участков. Для решения задачи регрессии также применялся метод CatBoost. Видно, что высокие содержания кальция в Ока-Урикском узле связаны с высоким содержанием натрия и фосфора, тогда как на картах корреляции (см. рис. 3) взаимосвязь между этими элементами практически отсутствует. Для Урда-Гарганского участка высокие содержания кальция связаны с высокими содержаниями Mg, Mn и Fe, что подтверждает присутствие силикатов марганца.

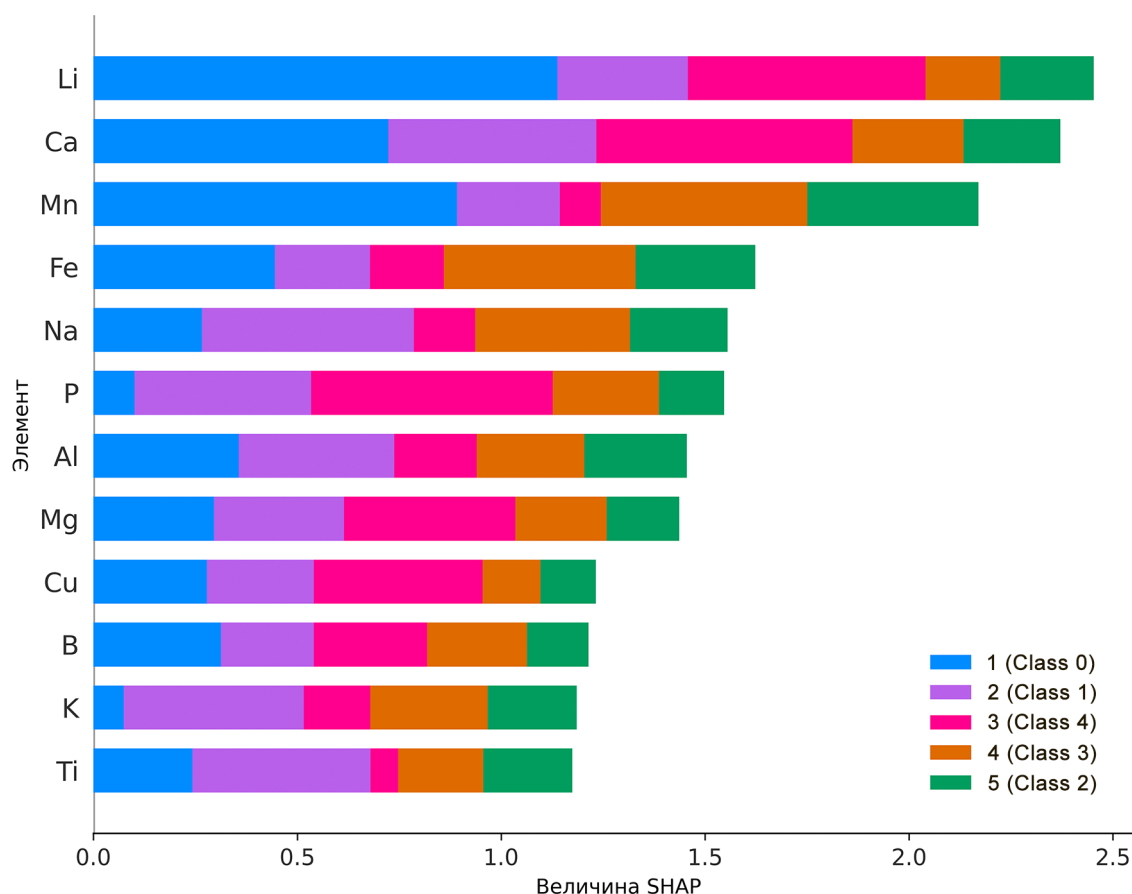


Рис. 7. SHAP диаграммы важности признаков, рассчитанные для метода CatBoost.

Целевая переменная – цвет кварцитов: 1 – белый, 2 – светло-серый, 3 – черный, 4 – темно-серый, 5 – серый.

Fig. 7. SHAP feature importance diagrams calculated for the CatBoost method.

The target variable is the color of quartzites: 1 – white, 2 – light gray, 3 – black, 4 – dark gray, 5 – gray.

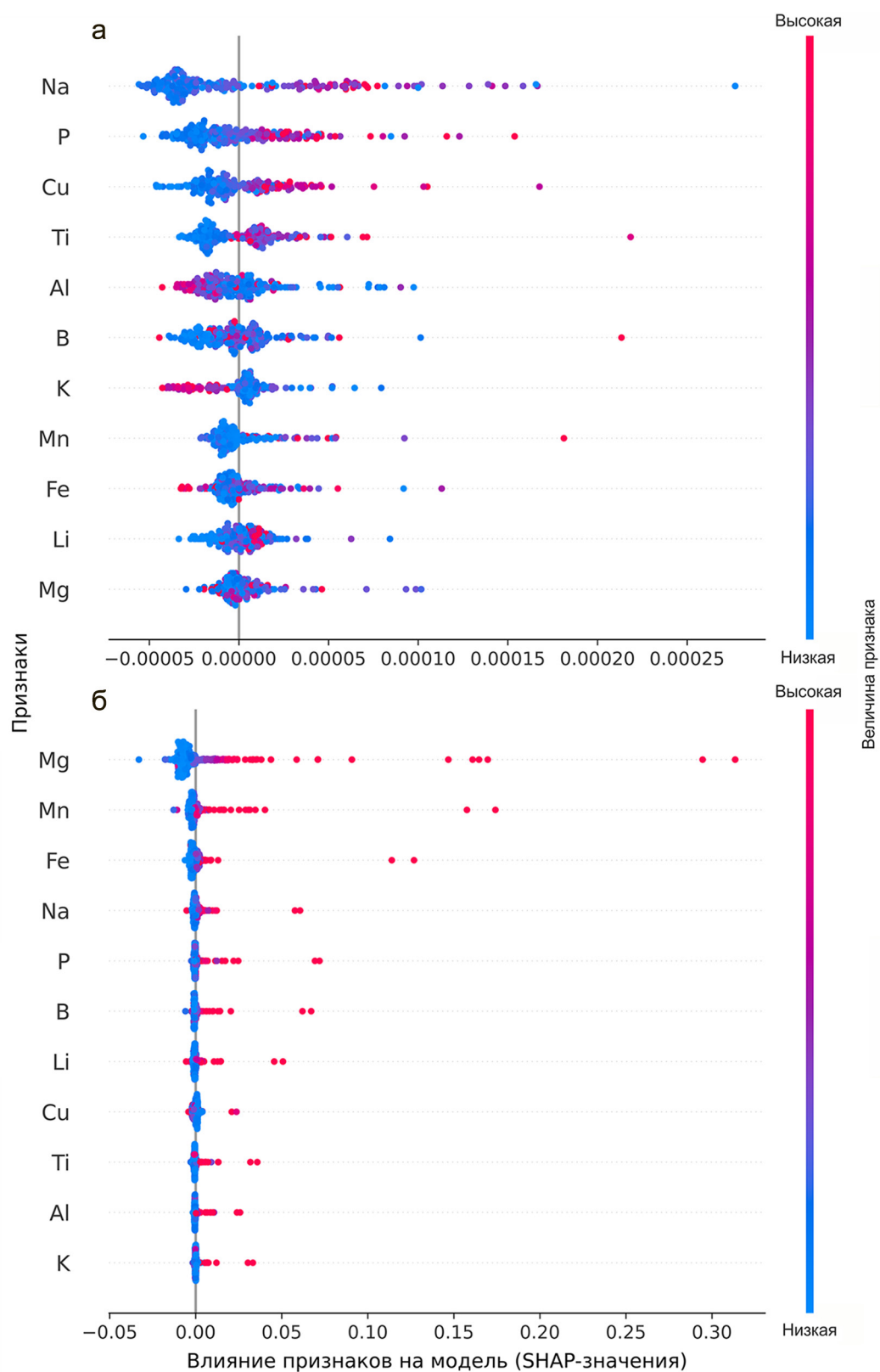


Рис. 8. SHAP диаграммы важности признаков, рассчитанные для метода CatBoost по содержанию Ca в кварцитах с Ока-Урикского (а) и Урда-Гарганского (б) участков.

Fig. 8. SHAP diagrams of feature importance calculated for the CatBoost method for the Ca content in quartzites from the Oka-Urisky (a) and Urda-Gargansky (b) sites.

Проведенный анализ выявил, что содержание марганца служит основным диагностическим признаком для различения кварцитов Ока-Урикского и Урда-Гарганского участков. Выявленная геохимическая особенность – существенное различие в содержании марганца при сопоставимых концентрациях других элементов – представляет особый научный интерес. Известно, что марганец склонен к образованию комплексных соединений и соосаждению с другими элементами в гипергенных условиях (Юдович, 2013). В нашем случае можно предположить несколько возможных механизмов такой дифференциации: биогенную аккумуляцию (что согласуется с повышенным содержанием углерода в Урда-Гарганских кварцитах), особенности седиментогенеза или постседиментационные миграционные процессы.

Особого внимания заслуживает отсутствие значимых корреляций марганца с другими элементами. Этот феномен может объясняться как низкими абсолютными концентрациями Mn (эффект “маскировки”), так и уникальностью механизма его концентрации в изучаемых породах. Доминирующая роль марганца в дифференциации месторождений, выявленная обоими методами машинного обучения, требует уточнения его минеральных форм, которые могут варьироваться от родонита до марганецсодержащих глинистых минералов или гидротермальных образований. Полная интерпретация выявленной геохимической специфики требует дополнительных исследований с привлечением минералогических и изотопных методов, что выходит за рамки настоящей работы, но будет предметом наших дальнейших исследований.

ЗАКЛЮЧЕНИЕ

Проведенное исследование продемонстрировало эффективность комплексного подхода, сочетающего традиционные статистические методы и современные алгоритмы машинного обучения для анализа геохимических особенностей кварцитов. Применение непараметрических статистических тестов выявило значимые различия в распределении элементов между Ока-Урикским и Урда-Гарганским участками, подтвердив целесообразность последующего использования методов машинного обучения.

Методы машинного обучения, в частности алгоритмы LightGBM, CatBoost и Random Forest, показали высокую эффективность (точность до 97%) в задаче классификации кварцитов по месторождениям. Комбинированный анализ важности признаков с использованием SHAP-значений и встроенных метрик `feature_importances_` позволил не только выявить ключевые диагностические элементы (марганец), но и установить сложные взаимосвязи между различными химическими компонента-

ми. Особый интерес представляет обнаруженная противоположная направленность влияния алюминия/калия и марганца/кальция/фосфора, что нашло объяснение в различии минерального состава кварцитов разных участков.

Интересным является и то, что применение методов машинного обучения для классификации по цвету кварцитов дало результаты с точностью до 0.94. Однако влияние различных факторов на описание цвета требует дальнейших исследований, учитывающих дополнительные параметры. Сравнительный анализ важности признаков показал, что методы машинного обучения обеспечивают более глубокое понимание взаимосвязей и зависимостей между химическими элементами по сравнению с традиционными корреляционными методами. Это позволяет уточнить и дополнить существующие представления о геохимических процессах входящих в состав извлекаемых образцов.

Разработанный методический подход, включающий применение RobustScaler для нормализации данных и t-SNE для визуализации, показал свою эффективность для работы с геохимическими данными, характеризующимися нелинейными зависимостями и наличием выбросов. Полученные результаты не только подтверждают существующие представления о минералогических особенностях изучаемых кварцитов, но и открывают новые перспективы для понимания процессов их формирования. Дальнейшие исследования должны быть направлены на минералогическую верификацию установленных элементных ассоциаций и изучение механизмов концентрации ключевых элементов, в первую очередь марганца.

Таким образом, использование современного машинного обучения открывает новые горизонты для анализа химического состава и позволяет более точно интерпретировать данные, что особенно актуально для геохимических исследований

СПИСОК ЛИТЕРАТУРЫ

- Аюржанаева Д.Ц., Федоров А.М., Мазукабзов А.М., Непомнящих А.И., Очирова Э.А. Посохов В.Ф. (2020) Механизмы формирования химически чистых кварцитов Бурал-Сардыкского месторождения. *Геология и геофизика*, **61**(10), 1316-1330. <https://doi.org/10.15372/GiG2023128>
- Непомнящих А.И., Федоров А.М., Жабоедов А.П., Волкова М.Г. (2023) Высокочистые кварциты Восточного Саяна. *Геология и геофизика*, **64**(8), 1205-1215. <https://doi.org/10.15372/GiG2023128>
- Федоров А.М., Макрыгина В.А., Мазукабзов А.М., Непомнящих А.И., Аюржанаева Д.Ц., Волкова М.Г. (2021) Ресурсы кварцевого сырья Гарганской зоны Восточно-Саянского кварцитоносного района. *Георесурсы*, **23**(4), 96-106. <https://doi.org/10.18599/grs.2021.4.11>
- Юдович Я.Э., Кетрис М.П. (2013) Геохимия марганца в процессах гипергенеза: обзор. *Биосфера*, **5**(1) 21-36.

- Baak M., Koopman R., Snoek H., Klous S. (2020) A new correlation coefficient between categorical, ordinal and interval variables with Pearson characteristics. *Comput. Stat. Data Anal.*, 152, 107043. <https://doi.org/10.1016/j.csda.2020.107043>
- Bajpai J., Singh S. (2019) On orthogonal hypergeometric groups of degree five. *Transact. Amer. Math. Soc.*, **372**(11), 7541. <https://doi.org/10.1090/tran/7677>
- Breiman L. (2001) Random Forests. *Machine Learning*, **45**(1), 5-32. <https://doi.org/10.1023/A:1010933404324>
- Chawla N.V., Bowyer K.W., Hall L.O., Kegelmeyer W.P. (2002) SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intellig. Res.*, 321-357. <https://doi.org/10.1613/jair.953>
- Fedorov A.M., Makrygina V.A., Nepomnyashchikh A.I., Zhaboedov A.P., Parshin A.V., Posokhov V.F., Sokolnikova Yu.V. (2019) Geochemistry and petrology of superpure quartzites from East Sayan Mountains, Russia. *Acta Geochim.* **38**(1), 22-39. <https://doi.org/10.1007/s11631-018-0268-5>
- Götze J., Möckel R. (2012) Quartz: Deposits, Mineralogy and Analytics. *Spring. Geol.*, Ö Springer-Verlag Berlin Heidelberg.
- Hnila P., Frahm E., Gilibert A. et al. (2025) "Open Sourcing" Workflow and Machine Learning Approaches for Attributing Obsidian Artifacts to Their Volcanic Origins: A Feasibility Study from the South Caucasus. *J. Archaeol. Method. Theory*, **32**(28).
- Ke G., Meng Q., Finley Th., Wang T., Chen W., Ma W., Ye Q., Liu T. (2017) LightGBM: A Highly Efficient Gradient Boosting Decision Tree. Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, 4-9 December 2017, 314s9-3157. <https://dl.acm.org/doi/10.5555/3294996.3295074>
- Lundberg S.M., Lee S.-I. (2017) A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*, **30**.
- Manning C.D., Raghavan P., Schütze H. (2008) Introduction to Information Retrieval. *Cambridge University Press*, 234-265.
- Müller A., Wanvik J.E., Ihlen P.M. (2012) Petrological and Chemical Characterisation of High-Purity Quartz Deposits with Examples from Norway. (Eds J. Götze, R. Möckel). *Quartz: Deposits, Mineralogy and Analytics*. Springer Geology. Springer, Berlin, Heidelberg.
- Pitblado B.L., Dehler C., Neff H., Nelson S.T. (2008) Pilot study experiments sourcing quartzite, Gunnison Basin, Colorado. *Geoarchaeology*, **23**(6), 742-778. <https://doi.org/10.1002/gea.20240>
- Pitblado B.L., Cannon M.B., Neff H., Dehler C.M., Nelson S.T. (2013) LA-ICP-MS analysis of quartzite from the Upper Gunnison Basin, Colorado. *J. Archaeol. Sci.*, **40**(4), 2196-2216. <https://doi.org/10.1016/j.jas.2012.11.016>
- Prokhorenkova L., Gusev G., Vorobev A., Dorogush A.V., Gulin A. (2017) CatBoost: Unbiased boosting with categorical features. arXiv:1706.09516v5. <https://doi.org/10.48550/arXiv.1706.09516>
- Shah S.A., Shao Y., Zhang Y., Zhao H., Zhao L. (2022) Texture and Trace Element Geochemistry of Quartz: A Review. *Minerals*, **12**, 1042. <https://doi.org/10.3390/min12081042>
- van der Maaten L.J.P. (2014) Accelerating t-SNE using Tree-Based Algorithms. *J. Mach. Learn. Res.*, **15**(93), 3221-3245.
- Wang Y., Qiu K.-F., Müller A., Hou Z.-L., Zhu Z.-H., Yu H.-C. (2021) Machine learning prediction of quartz forming-environments. *J. Geophys. Res.: Solid Earth*, **126**, e2021JB021925. <https://doi.org/10.1029/2021JB021925>
- Yu H.F., Huang F.L., Lin C.J. (2011) Dual coordinate descent methods for logistic regression and maximum entropy models. *Mach. Learn.*, **85**, 41-75. <https://doi.org/10.1007/s10994-010-5221-8>
- Zhu G.-D., Niu Y.-Y., Liao S.-B., Ruan L., Zhang X.-H. (2023) Discrimination of Quartz Genesis Based on Explainable Machine Learning. *Minerals*, **13**, 997. <https://doi.org/10.3390/min13080997>

REFERENCES

- Ayurzhanyeva D.Ts., Fedorov A.M., Mazukabzov A.M., Nepomnyashchikh A.I., Ochirova E.A., Posokhov V.F. (2020) Mechanisms of High-Purity Quartzite Formation at the Bural-Sardyk Deposit (Russia). *Geol. Geophys.*, **61**(10), 1316-1330. <https://doi.org/10.15372/GiG2023128>
- Baak M., Koopman R., Snoek H., Klous S. (2020) A new correlation coefficient between categorical, ordinal and interval variables with Pearson characteristics. *Comput. Stat. Data Anal.*, 152, 107043. <https://doi.org/10.1016/j.csda.2020.107043>
- Bajpai J., Singh S. (2019) On orthogonal hypergeometric groups of degree five. *Transact. Amer. Math. Soc.*, **372**(11), 7541. <https://doi.org/10.1090/tran/7677>
- Breiman L. (2001) Random Forests. *Machine Learning*, **45**(1), 5-32. <https://doi.org/10.1023/A:1010933404324>
- Chawla N.V., Bowyer K.W., Hall L.O., Kegelmeyer W.P. (2002) SMOTE: Synthetic minority over-sampling technique. *J. Artif. Intellig. Res.*, 321-357. <https://doi.org/10.1613/jair.953>
- Fedorov A.M., Makrygina V.A., Mazukabzov A.M., Nepomnyashchikh A.I., Ayurzhanyeva D.Ts., Volkova M.G. (2021) Resources of quartz raw materials, Gargan block, East Sayan quartzite-bearing area. *Georesources*, **23**(4), 96-106. <https://doi.org/10.18599/grs.2021.4.11>
- Fedorov A.M., Makrygina V.A., Nepomnyashchikh A.I., Zhaboedov A.P., Parshin A.V., Posokhov V.F., Sokolnikova Yu.V. (2019) Geochemistry and petrology of superpure quartzites from East Sayan Mountains, Russia. *Acta Geochim.* **38**(1), 22-39. <https://doi.org/10.1007/s11631-018-0268-5>
- Götze J., Möckel R. (2012) Quartz: Deposits, Mineralogy and Analytics. *Spring. Geol.*, Ö Springer-Verlag Berlin Heidelberg.
- Hnila P., Frahm E., Gilibert A. et al. (2025) "Open Sourcing" Workflow and Machine Learning Approaches for Attributing Obsidian Artifacts to Their Volcanic Origins: A Feasibility Study from the South Caucasus. *J. Archaeol. Method. Theory*, **32**(28).
- Ke G., Meng Q., Finley Th., Wang T., Chen W., Ma W., Ye Q., Liu T. (2017) LightGBM: A Highly Efficient Gradient Boosting Decision Tree. Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, 4-9 December 2017, 314s9-3157. <https://dl.acm.org/doi/10.5555/3294996.3295074>
- Lundberg S.M., Lee S.-I. (2017) A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*, **30**.

- Manning C.D., Raghavan P., Schütze H. (2008) Introduction to Information Retrieval. *Cambridge University Press*, 234-265.
- Müller A., Wanvik J.E., Ihlen P.M. (2012) Petrological and Chemical Characterisation of High-Purity Quartz Deposits with Examples from Norway. (Eds J. Götze, R. Möckel). *Quartz: Deposits, Mineralogy and Analytics*. Springer Geology. Springer, Berlin, Heidelberg.
- Nepomnyashchikh A.I., Fedorov A.M., Zhaboev A.P., Volkova M.G. (2023) High-Purity Quartzite from East Sayan. *Geol. Geophys.*, **64**(8), 1205-1215. <https://doi.org/10.15372/GiG2023128>
- Pitblado B.L., Cannon M.B., Neff H., Dehler C.M., Nelson S.T. (2013) LA-ICP-MS analysis of quartzite from the Upper Gunnison Basin, Colorado. *J. Archaeol. Sci.*, **40**(4), 2196-2216. <https://doi.org/10.1016/j.jas.2012.11.016>
- Pitblado B.L., Dehler C., Neff H., Nelson S.T. (2008) Pilot study experiments sourcing quartzite, Gunnison Basin, Colorado. *Geoarchaeology*, **23**(6), 742-778. <https://doi.org/10.1002/gea.20240>
- Prokhorenkova L., Gusev G., Vorobev A., Dorogush A.V., Gulín A. (2017) CatBoost: Unbiased boosting with categorical features. [arXiv:1706.09516v5](https://arxiv.org/abs/1706.09516). <https://doi.org/10.48550/arXiv.1706.09516>
- Shah S.A., Shao Y., Zhang Y., Zhao H., Zhao L. (2022) Texture and Trace Element Geochemistry of Quartz: A Review. *Minerals*, **12**, 1042. <https://doi.org/10.3390/min12081042>
- van der Maaten L.J.P. (2014) Accelerating t-SNE using Tree-Based Algorithms. *J. Mach. Learn. Res.*, **15**(93), 3221-3245.
- Wang Y., Qiu K.-F., Müller A., Hou Z.-L., Zhu Z.-H., Yu H.-C. (2021) Machine learning prediction of quartz forming-environments. *J. Geophys. Res.: Solid Earth*, **126**, e2021JB021925. <https://doi.org/10.1029/2021JB021925>
- Yu H.F., Huang F.L., Lin C.J. (2011) Dual coordinate descent methods for logistic regression and maximum entropy models. *Mach. Learn.*, **85**, 41-75. <https://doi.org/10.1007/s10994-010-5221-8>
- Yudovich Ya.E., Ketris M.P. (2013) Manganese geochemistry in hypergenesis processes: A review. *Biosphere*, **5**(1), 21-36.
- Zhu G.-D., Niu Y.-Y., Liao S.-B., Ruan L., Zhang X.-H. (2023) Discrimination of Quartz Genesis Based on Explainable Machine Learning. *Minerals*, **13**, 997. <https://doi.org/10.3390/min13080997>