Санкт–Петербургский государственный университет

САЛИМОВ Тимур Альфредович

Выпускная квалификационная работа

Автоматическая типизация полноразмерного керна

Уровень образования: бакалавриат
Направление 02.03.02 «Фундаментальная информатика и информационные технологии»
Основная образовательная программа СВ.5003.2016 «Программирование и информационные технологии»
Профиль «Автоматизация научных исследований»

Научный руководитель:
доцент, кафедра компьютерных технологий и систем,
к.ф-м.н. Погожев Сергей Владимирович

Рецензент:
ООО «Газпромнефть-Цифровые решения»
Горбунов Владислав Игоревич

Санкт-Петербург
2020 г.

Содержание

Введение

Эффективное ведение геологоразведочных работ при поиске и разведке скоплений углеводородов и последующей разработке залежей невозможно без детального и своевременного исследования извлеченного при бурении керна, который является главным носителем реальной информации о недрах земли. Керн представляет собой цилиндрический образец несколько дециметров в диаметре и несколько метров в длину. Обычно он состоит из 3-5 широких слоёв породы, чередующихся узкими прослойками (слойками) включений. Будучи добытым из коллектора углеводородов, в состав керна могут входить различные нефтеносные осадочные породы, например, несцементированный песок, известняк, доломит, песчаник, алевролит, а также слои глины, образующие природные границы резервуара. В первую очередь керн распиливают на несколько частей, полируют и целиком фотографируют при дневном и ультрафиолетовом (УФ) свете. Последнее делают для визуальной идентификации нефти: при дневном свете нефтяные включения (особенно для образцов с крупными порами/трещинами и высокой насыщенностью) видны как тёмные маслянистые пятна на поверхности; в УФ свете нефть обычно люминисцирует бело-синим на фото, а иногда коричнево-жёлтым в зависимости от её сорта. Даже если в образце нефть сразу не обнаруживают, для него по-прежнему идентифицируют фации (слои, имеющие общие характеристики), определяют степень раздробленности пластов и идентифицируют литотипы пород, слагающих керн. Эта информация, впоследствии используется для построения модели месторождения.

Создание геологических моделей нефтяных и газовых месторождений и прогнозирование распространения ловушек углеводородов должно опираться на результаты исследования керна, позволяющие получить достоверные сведения о строении нефтегазовых комплексов, о составе отложений, о характере насыщения и других свойствах пород-коллекторов. С изучения керна начинается поиск, разведка, доразведка, а в дальнейшем и разработка любого месторождения. Бурение и всестороннее исследование керна практически единственные методы, которые позволяют полу-

чить достоверную информацию о свойствах и составе веществ на больших глубинах и обеспечивают объективную проверку и интерпретацию дистанционных исследований.

На данный момент один из методов первичного анализа заключается в описании основных характеристик керна по фотографиям в дневном свете и после люминесцентно-битуминологического анализа [1]. Необходимо определить тип породы, нефтенасыщенность, карбонатность и разрушенность керна.

Постановка задачи

Цель работы

Целью работы является создание прототипа для автоматической разметки фотографий керна с применением методов машинного обучения, который автоматизирует процесс описания керна по фотографиям и сократит необходимое для анализа время. Такой программный комплекс потенциально позволит:

- уменьшить общую длительность цикла геологоразведочных работ;
- снизить расходы на оплату работ подрядной сервисной организации;
- повысить точность определения литотипов за счёт снижения роли человеческого фактора;
- систематизировать в едином формате полученную информацию о текстурно-структурных особенностях пород.

Ожидается, что в целевом виде такое ПО сократит время литологического анализа керна без необходимости личного выезда экспертов в кернохранилище, что также занимает дополнительное время при анализе образцов.

Задачи работы

Для достижения цели были поставлены следующие задачи:

- исследовать имеющуюся базу данных фотографий керна;
- разработать гипотезы для применения распознавания образов и предиктивной аналитики;
- проверить работоспособность прототипа;
- разработать программный комплекс для обогащения имеющейся базы данных фотографий керна;

- провести исследования для подтверждения увеличения качества автоматической типизации фотографий керна при обогащении данных;
- провести тестирование разработанного программного комплекса и оценку качества разработанного подхода.

Обзор литературы

Для анализа фотографий керна и их изучения в работе [2] рассматриваются методы распознавания образов и компьютерной обработки, позволяющие извлечь дополнительную информацию из изображений. В работе уделено особое внимание аналитическим методам и их модификациям, которые свойственны данной предметной области. В работе с изображениями керна есть свои трудности и аспекты, на которые следуют обращать внимание, поэтому при отборе атрибутов для описания изображения был сделан упор на рекомендуемые методы из данной работы. Помимо анализа изображений стояла также задача по классификации как самого изображения, так и отдельных областей фотографий. Для решения этой задачи были рассмотрены следующие алгоритмы классификации:

- Random Forest Classifier – алгоритм машинного обучения, основанный на применении множества деревьев решений. Это ансамблевый метод, основанный на идее бэкинга. Алгоритм использует усреднение для повышения точности прогнозирования[3] и является устойчивым к шумам, что необходимо было учитывать при работе с исходной разметкой;

- Extra Tree Classifier – данный алгоритм имеет такой же принцип, как и описанный выше Random Forest Classifier, но в качестве базовой модели использует рандомизированное решающее дерево [4], что позволяет сократить время обучения и использовать меньшие вычислительные ресурсы;

- SVM Classifier – метод опорных векторов для классификации. Исходные вектора атрибутов переводятся в пространство более высокой размерности, а затем в этом пространстве вычисляется разделяющая гиперплоскость с максимальным расстоянием от объектов всех представленных классов [5];

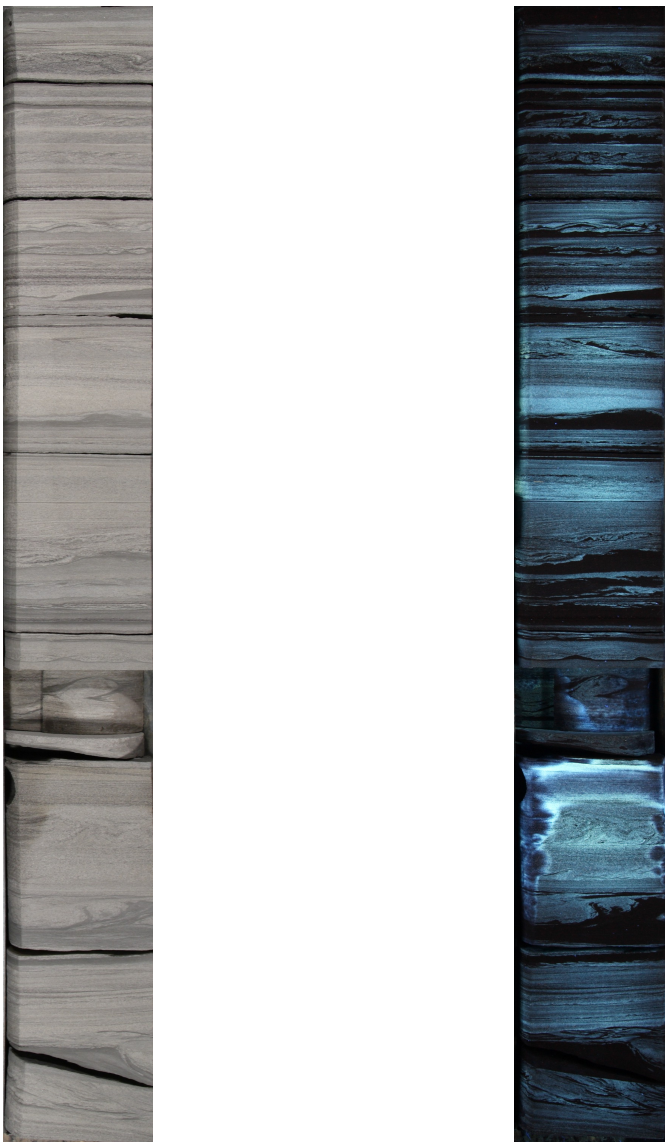
Глава 1. Анализ фотографий керна

Из кернохранилища были получены фотографии керна с различных месторождений. По дате и условиям съемки, условиям хранения и транспортировке фотографии сильно различались, что имеет значительное влияние при создании математической модели. В следующих разделах будут рассмотрены имеющиеся данные, их распределение, предметное описание и выделенные особенности.

1.1 Описание исходных данных

Данные представляли из себя пары фотографий, пример которых приведен на рис. 1. Осадочные толщи имеют слоистое, часто периодическое, строение и представляют многократное и разномасштабное чередование пород [6]. Поэтому при осмотре и описании керновых колонок, прежде всего, выделяются слои – геологические тела, имеющие существенно однородный литологический состав, например одинаковую окраску, обладающие ясно выраженными подошвой и кровлей и значительной толщиной. Внутри слоев выделяются слойки – первичные элементы слоистости, «обособленные в теле слоя элементы более мелкого масштаба, имеющие визуально различимые границы ограничения». Слойки могут быть сгруппированы в серии. В случае частого ритмичного переслаивания слойков допускается однократное, подробное описание особенностей каждой из чередующихся разностей пород с последующим указанием лишь расположения и толщины повторяющихся слоев и тех изменений, какие отмечаются по отношению к уже описанным.

Помимо фотографий были предоставлены табличные данные с описанием по всем четырем параметрам для каждой пары фотографий. Однако в предоставленной табличной разметке для каждой фотографии в соответствие ставится описание содержащее лишь один тип породы, определенный тип разрушенности, карбонатности и нефтенасыщенности. В результате сбора данных была сформирована выборка из 17740 фотографий и табличного описания этих фотографий. Информация о месторождении и скважины, откуда бралась проба были объединены и закодированы. Эта



(a) Дневной свет **(b)** Ультрафиолетовое излучение

Рис. 1: Пример данных

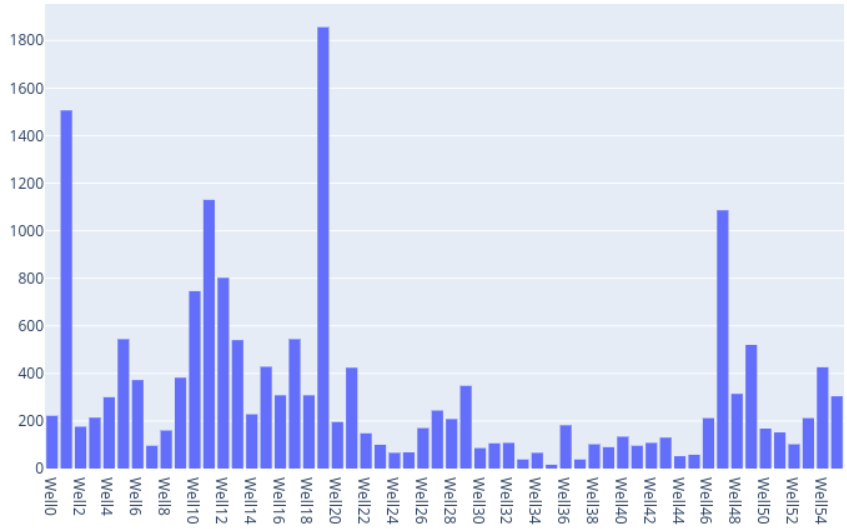


Рис. 2: Распределение по месторождениям

информация является важным фактором, так как на месторождениях может быть использовано различное оборудование и могут разниться условия съемки. Распределение по месторождениям (рис. 2) учитывалось при формировании обучающей выборки.

Как видно из рисунка 3, имеется большее количество типов пород в исходных данных. Однако для большинства из них предоставлено малое количество примеров для обучения, что существенно влияет на качество работы моделей машинного обучения. Наиболее крупные классы в начальной выборке: песчаник, алевролит, аргиллит, а также переслой песчаника, алевролита и глин. Однако переслой в себе сочетает несколько разных пород. Это может отразиться на результате обучения, поэтому переслои не использовались для обучения моделей.

Как показано на графиках распределения по классам нефтенасыщенности и карбонатности (рис. 4 и 5) большинство экземпляров является не нефтенасыщенными и не карбонатными. Поэтому для обучения все степени нефтенасыщенности объединялись в один класс “нефтенасыщенный”. Таким же образом обрабатывались данные по карбонатности.

Исходя из графика распределения по классам разрушенности (рис. 6) можно также сделать вывод что, и эта задача является мультиклассовой классификацией. Однако, как и в случае с насыщенностью и типами пород, большинство примеров из имеющихся данных являются не разрушенными.

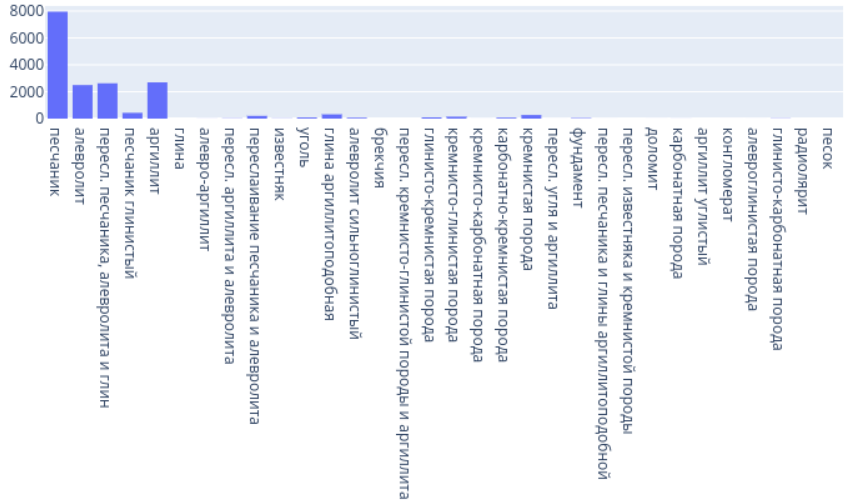


Рис. 3: Распределение по типам пород

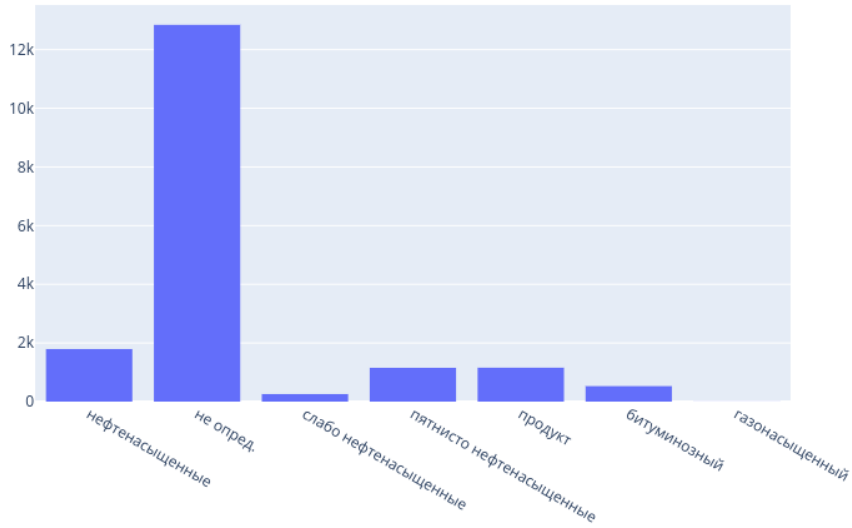


Рис. 4: Распределение по нефтенасыщенности

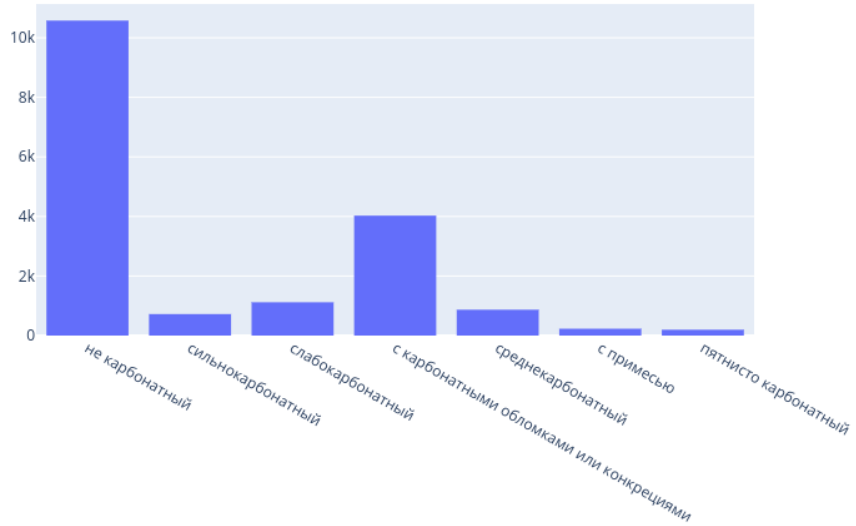


Рис. 5: Распределение по карбонатности

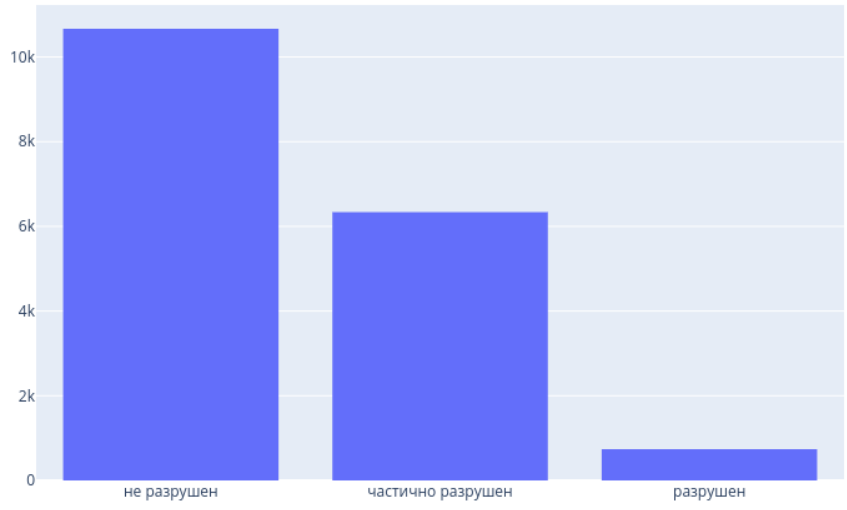


Рис. 6: Распределение по разрушенности

Поэтому для обучения все степени разрушенности(частично разрушен и разрушен) объединялись в один класс “разрушен”, следовательно, задача также сводилась к бинарной классификации.

1.2 Формирование обучающей выборки

В первую очередь выбирались фотографии, площадь которых была больше 1 млн. пикселей, также при помощи ручной разметки отбрасывались размытые фотографии и те, что содержали посторонние элементы. Например, на некоторых фотографиях были видны ящики для хранения керна, таблички с подписями от сотрудников и прочие элементы, не имеющие отношение к задаче. Однако необходимо было оставить достаточное количество примеров, для применения методов машинного обучения. Также были трудности из-за общего количества примеров, что не позволяло провести полную ручную разметку. Оставшиеся фотографии формировали обучающую выборку, с учетом распределения по месторождениям.

При использовании моделей машинного обучения, существует вероятность, что модель будет игнорировать классы, которые составляют меньшинство. Работа с несбалансированными данными имеет свои трудности и следует прибегать к такому типу машинного обучения лишь при необходимости. Так как в рамках этого проекта не было определено, что один класс должен предсказывать с большей вероятностью чем другие, или что необходимо учитывать частоту появления породы из обучающей выборки в модели, была проведена балансировка данных перед обучением. Так как в представленной выборке из побочной информации были лишь месторождения и скважины, балансировка производилось с учётом разного географического расположения скважин, из которых бралась проба. Одной из сложностей при обработке выборки было то, что данные были получены при съемке на разное оборудование и при разных условиях, то есть необходимо было заложить в модель возможность предсказывать тип породы по фотографиям с учётом различных условий съемки представленных в выборке. Процесс балансировки проходил следующим образом:

1. Выбирались фотографии нужного типа (определенные типы пород,

насыщенности, разрушенности);

2. Затем вычислялась гистограмма распределения выбранных фото по месторождениям и скважинам. Ожидаемо фотографии были распределены неравномерно;

3. После определялось максимальное количество примеров для каждого класса и выбирался минимум;

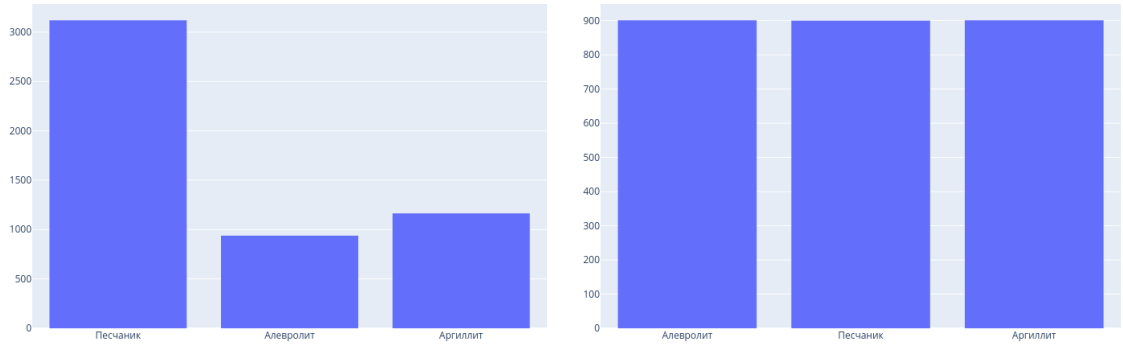
4. Равномерно выбирались примеры из каждой скважины, пока выборка не наполнялась до необходимого размера. Для классов с большим количеством примеров происходило перемешивание;

Таким образом в обучающей выборке обязательно были экземпляры из различных скважин. На рисунках 7, 8 и 9 представлено распределение изображений в обучающих выборках до и после балансировки.

При формировании выборки по типу породы, были выбраны три породы, которые чаще всего встречались в имеющихся данных (песчаник, алевролит и аргиллит). В итоге для обучения использовались по 900 экземпляров каждого класса, то есть 2700 фотографий суммарно.

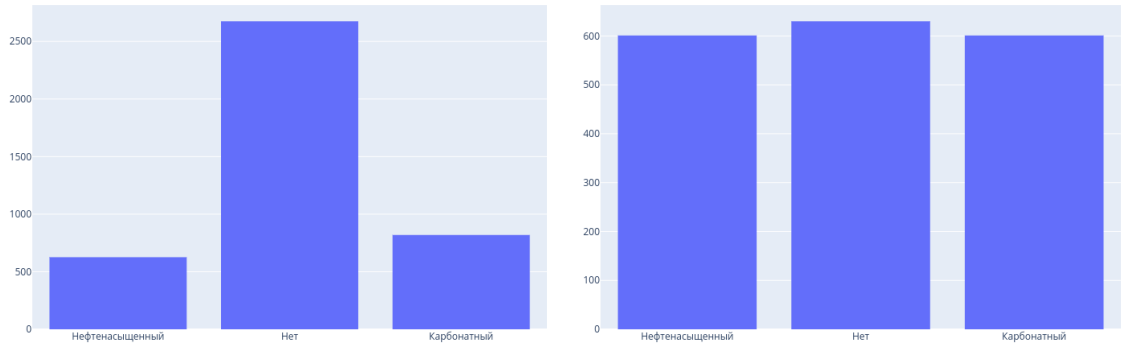
Специалистами по предметной области было отмечено, что одновременно в керне не может присутствовать и карбонатность и нефтенасыщенность. Небольшой процент таких случаев был в исходных данных, они были отброшены. Так же разбитые на множественные типы классы "карбонатность"и "нефтенасыщенность"были сокращены до бинарной классификации, а позже из-за наличия корреляции (если нефтенасыщенный, то не карбонатный) были объединены в один класс. Таким образом были отобраны по 600 экземпляров для каждого класса, что суммарно дало 1800 фотографий для обучения.

Для определения разрушенности удалось выбрать по 2900 примеров разрушенных и не разрушенных столбов керна, то есть суммарно исследование проводилось на 5800 фотографиях.



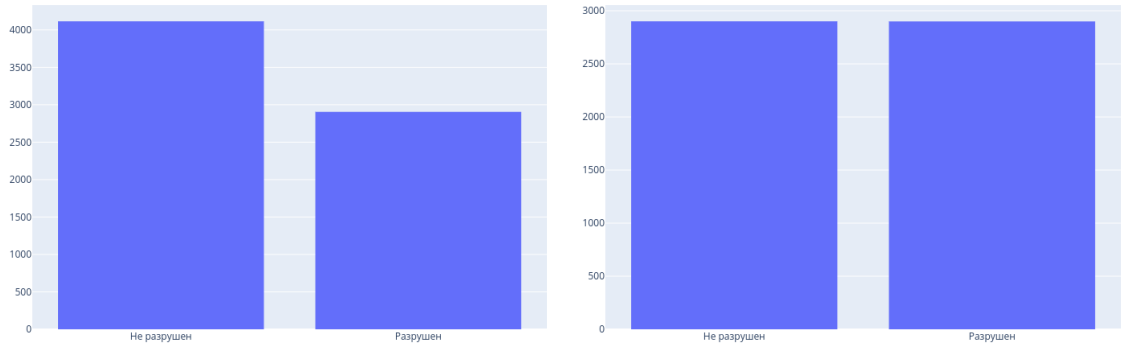
(a) До балансировки **(b)** После балансировки

Рис. 7: Распределение по типам пород



(a) До балансировки **(b)** После балансировки

Рис. 8: Распределение по нефтенасыщенности и карбонатности



(a) До балансировки **(b)** После балансировки

Рис. 9: Распределение по разрушенности

1.3 Выделение атрибутов для обучения

1.3.1 Сегментация изображений керна

На изображении могут присутствовать различные типы породы, переслаивания и трещины. Поэтому первым этапом для формирования табличных данных на основе фотографий стала задача сегментации изображения. Использовались классические подходы из теории обработки изоб-

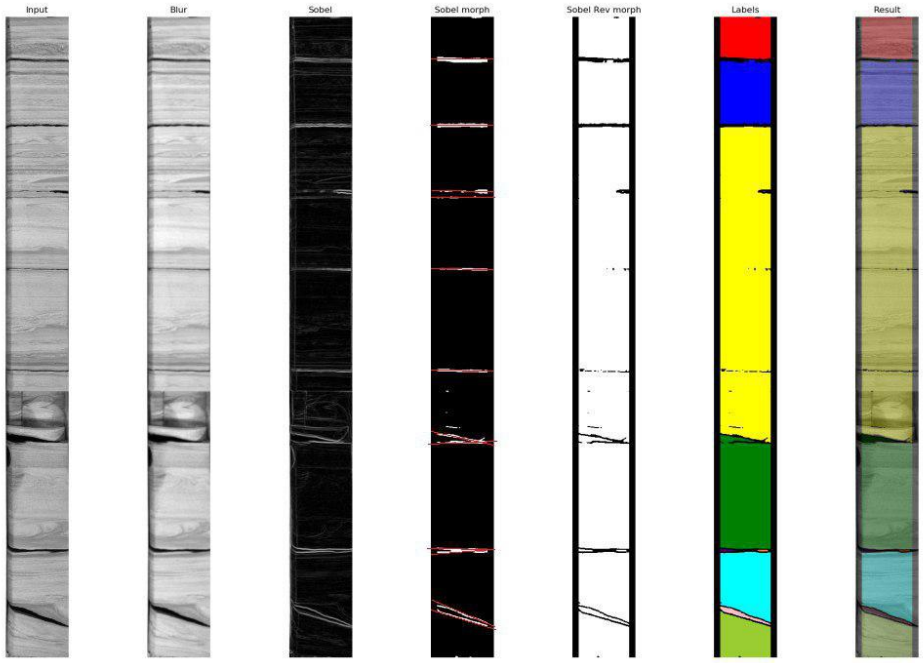


Рис. 10: Пример сегментации изображения

ражений: бинаризация по порогу, поиск границ, методы заливки, для объединения однотипных областей. Эта задача до сих пор не была решена полностью, так как фотографии керна из имеющихся данных были получены при разных условиях и при помощи различного оборудования.

1.3.2 Дескрипторы для сегментов

После сегментации для каждой области на фотографии применялись дескрипторы, позволяющие привести матричное представление разных сегментов к унифицированному табличному описанию. Описать изображение

можно по текстуре, форме и интенсивности. Были выбраны дескрипторы, которые покрывают все эти области, так же большая часть дескрипторов была глобальной, чтобы не привязываться к небольшим локальным особенностям, потому что сам по себе этот тип породы осадочный, составной, поэтому стоит рассматривать области в первую очередь по глобальным параметрам [7].

1.3.3 Матрица смежности

Матрица частот пар пикселей определенной яркости, расположенных на изображении определенным образом относительно друг друга.

$$C_{\Delta x,\Delta y}(i,j) = \sum_x \sum_y \begin{cases} 1, & \text{if } I(x,y) = i \text{ and } I(x+\Delta x, y+\Delta y) = j \\ 0, & \text{otherwise} \end{cases}$$

где i и j значение пикселей; x и y расположение на изображении I; смещение $(\Delta x, \Delta y)$ определяет расположение, относительно которого рассчитывается матрица; и I(x,y) указывает значение пикселя в точке (x,y).

На основании этой матрицы вычислялись дополнительные характеристики:

1. Энергия - минимален, когда все элементы равны:

$$Energy = \sum_i \sum_j C^2(i,j)$$

2. Энтропия - мера хаотичности, максимален, когда все элементы равны:

$$Entropy = -\sum_i \sum_j C(i.j) \log_2 C(i,j)$$

3. Контраст - мал, когда большие элементы вблизи главной диагонали:

$$Contrast = \sum_i \sum_j (i-j)^2 C(i,j)$$

4. Обратный разностный момент - мал, когда большие элементы далеки от главной диагонали:

$$InverseDifferenceMoment = \sum_i \sum_j \frac{C(i,j)}{1+(i-j)^2}$$

1.3.4 Локальные бинарные паттерны

Локальные бинарные паттерны (англ. Local Binary Patterns, LBP) — простой оператор, используемый для классификации текстур в компьютерном зрении. ЛБШ представляет собой описание окрестности пикселя изображения в двоичной форме. Оператор ЛБШ, который применяется к пикселю изображения, использует восемь пикселей окрестности, принимая центральный пиксель в качестве порога. Пиксели, которые имеют значения больше, чем центральный пиксель (или равное ему), принимают значения 1, а те, которые меньше центрального, принимают значения 0. Таким образом получается восьмиразрядный бинарный код, который описывает окрестность пикселя. Затем вычислялась гистограмма распределения полученных восьмиразрядных кодов и использовалась в качестве вектора атрибутов.

1.3.5 Глобальное описание сегмента

Для сегмента вычислялась максимальная, средняя и минимальная интенсивности в области, а также моменты интенсивностей пикселей изображения. Подбирались моменты, которые были бы инварианты к переносу, повороту и растяжению.

1.4 Обогащение данных

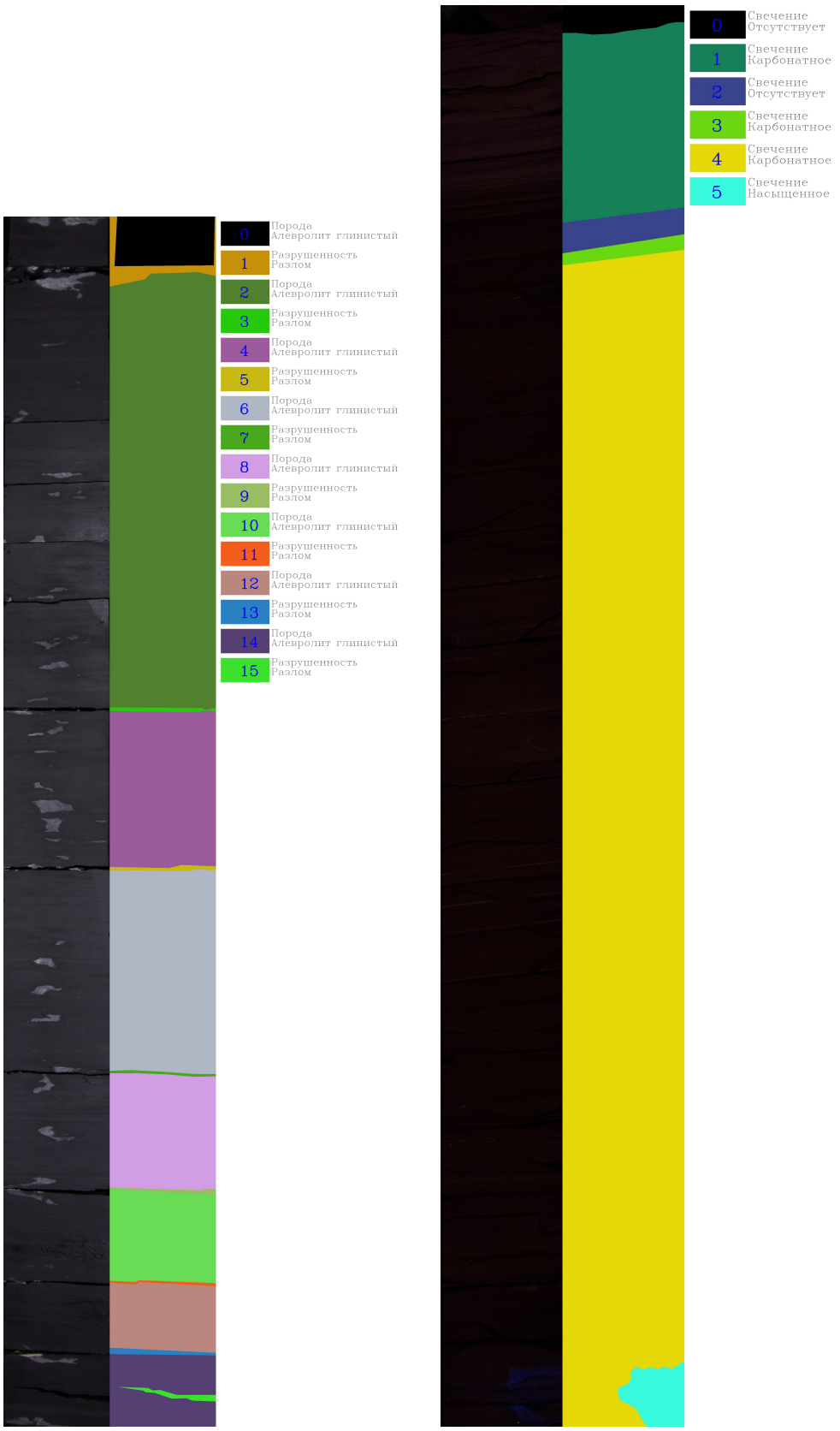
По результатам проведенных исследований, которые описаны в следующих разделах, было принято решение обогатить имеющиеся данные с помощью экспертов предметной области - литологов. Основная проблема заключалась в следующем: для одной фотографии, представляющей собой один метр керна в большинстве случаев, имеется описание лишь преоб-

ладающего типа пород, свечения или же общая степень разрушения без локализации описываемой области и детального рассмотрения, различающихся между собой, областей фотографии. Таким образом при использовании исходной разметки в процессе обучения не удавалось четко отделить различные по целевому показателю области, так как в обучающие выборки попадали ложные примеры.

Для детализации описания фотографий керна было принято решение разработать программный продукт, позволяющий при помощи специалистов провести детальное описание фотографий керна. Описание разработанного приложения будет приведено в следующих разделах.

После двух месяцев работы приложения удалось детализировать разметку для более чем 12000 фотографий керна. Разработанный программный продукт позволил получить матрицу сегментации, которая строго локализировала описываемую область, а также более детальное описание областей фотографий благодаря расширенному списку параметров для одного сегмента 11.

Собранные данные позволят не только улучшить результаты предиктивной аналитики для типизации фотографий керна, но также провести исследования связанные с сегментацией изображения керна и выявления особенностей экспертной разметки фотографий.



(a) Описание фотографии в дневном свете

(b) Описание фотографии в ультрафиолетовом свете

Рис. 11: Примеры экспертной разметки фотографий керна

Глава 2. Разработка программного комплекса

2.1 Организация программного окружения для проведения исследований

При разработке и организации программного окружения для проведения исследований основной упор был выполнен на воспроизводимость полученных результатов. В результате были выбраны следующие программные продукты и технологии:

- Язык программирования Python 3 – основной язык для реализации алгоритмов обработки и анализа данных, а также применения моделей машинного обучения [8];
- Snakemake – система управления рабочим процессом, позволяющая создавать воспроизводимые и масштабируемые исследования с использованием языка Python 3. Настроенный рабочий процесс можно легко масштабировать для серверной, кластерной, сеточной и облачной среды выполнения, без необходимости кардинальных изменений рабочего процесса;
- MinIO – это сервер облачного хранилища, совместимый с Amazon S3. Как хранилище объектов, MinIO может хранить неструктурированные данные, такие как фотографии;
- Scikit-Learn – библиотека с реализованными моделями машинного обучения и унифицированным интерфейсом организации рабочего процесса;
- Scikit-Image – библиотека обработки изображений, включающая в себя алгоритмы сегментации, геометрических преобразований, обнаружения признаков и прочие алгоритмы для работы с изображениями и их анализа;
- Pandas – библиотека для работы с данными в табличном представлении, а также с файлами формата CSV и Microsoft Excel;

1. Download data – загрузка описания фотографий из кернохранилища;
2. Download images – загрузка исходных фотографий керна;
3. Download matrixes – загрузка экспертной разметки фотографий;
4. Download slice – загрузка описания экспертной разметки;
5. Make sample – список фотографий для обучения моделей;
6. Make features – формирование набора сегментов и их признаков;
7. Make dataset – формирование dataset`а для обучения;
8. Split dataset – разделение dataset`а на обучающую и тестовую выборки;
9. Train model – этап обучения моделей;
10. General – формирование результатов провидения эксперимента;

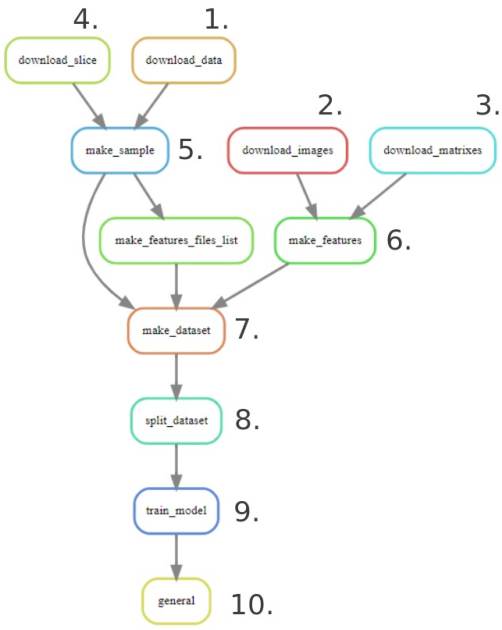


Рис. 12: Схема процесса работы для получения итоговых моделей

В качестве облачного хранилища использовалось MinIO, стоит отметить что одним из преимуществ данного программного обеспечения является его нетребовательность к вычислительным ресурсам, а также простота и гибкость в настройке. На сервере хранились как исходные фотографии полученные из кернохранилища, так и собранная экспертная разметка и обработанные данные для создания моделей машинного обучения. Основная идея Snakemake в том, что рабочий процесс можно представить как последовательное преобразование файлов и получения на их основе новых файлов. На первом этапе сформированного рабочего процесса происходила автоматическая загрузка данных из облачного хранилища, затем для загруженных фотографий выполнялась сегментация, либо использовалась экспертная разметка в зависимости от выполняемого исследования. После для выделенных на фотографиях сегментах вычислялись дескрипторы для приведения данных к табличному виду. Следующим шагом происходило создание моделей машинного обучения, а также выполнялось их тестирование на сформированной выборке фотографий. Схематично разработанный процесс формирования обученных моделей для предиктивного анализа фотографий керна представлен на рисунке 12.

2.2 Разработка приложения для экспертной разметки фотографий керна

- Django – фреймворк для веб-приложений на языке Python, позволяющий разрабатывать модульное приложение, с возможностью интеграции с другими сервисами и продуктами;
- React – JavaScript-библиотека для создания пользовательского интерфейса веб-приложения;
- PostgreSQL – объектно-реляционная система управления базами данных;

Было принято решение о разработке веб-приложения, что позволило не привязываться к определенной платформе и в кротчайшие сроки приступить к работе с экспертами. Для каждого эксперта был сформирован список заданий, который представлял из себя набор изображений керна как в дневном свете, так и в ультрафиолете. После авторизации в системе пользователю отправлялось одно из его заданий и предоставлялись инструменты для ручной разметки фотографии.Пользователь разделяет кривыми линиями области различных слоев керна, обозначая граничные переходы (рис. 13) и получает список сегментов для присваивания им значений (рис. 14). Далее пользователь переходит к разметке сегментов, определенных в промежутках между разными видами граничных переходов. Для выбранного типа сегмента пользователь может выбрать соответствующий набор параметров, пример которых приведен на рис. 15.

В начале с помощью экспертов предметной области была составлена схема пользовательского опроса. В итоговом варианте опроса оставались только целевые параметры для задачи распознавания, а также вспомогательные, которые могут повлиять на предсказание целевого показателя, а также поддаются экспертной разметке лишь по фотографии, без использования других методов анализа.

Была собрана команда из десяти экспертов литологов, которая на протяжении двух месяцев использовала разработанное веб-приложение для

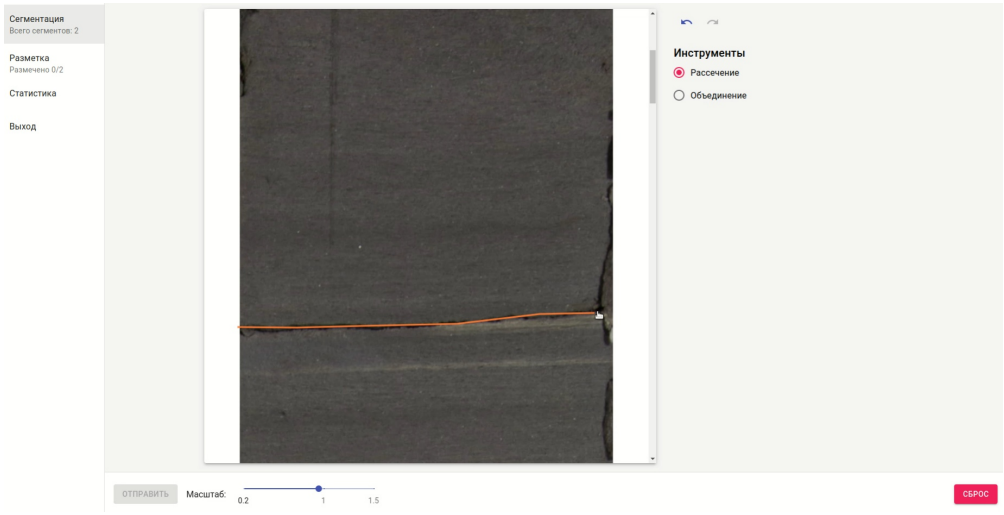


Рис. 13: Пример работы ручной сегментации

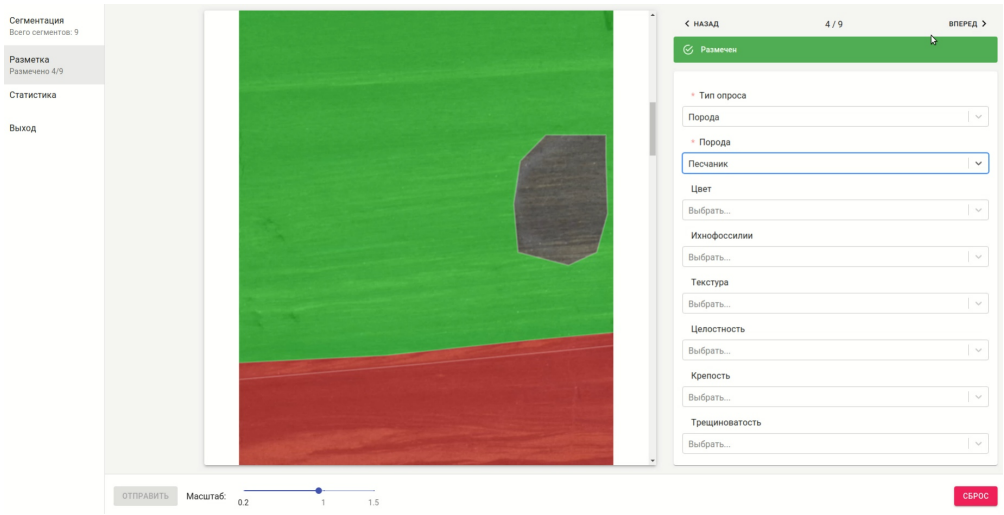


Рис. 14: Пример работы с заполнением опроса о фотографии

обогащения исходных данных. Удалось получить экспертную разметку для 12489 фотографий керна.

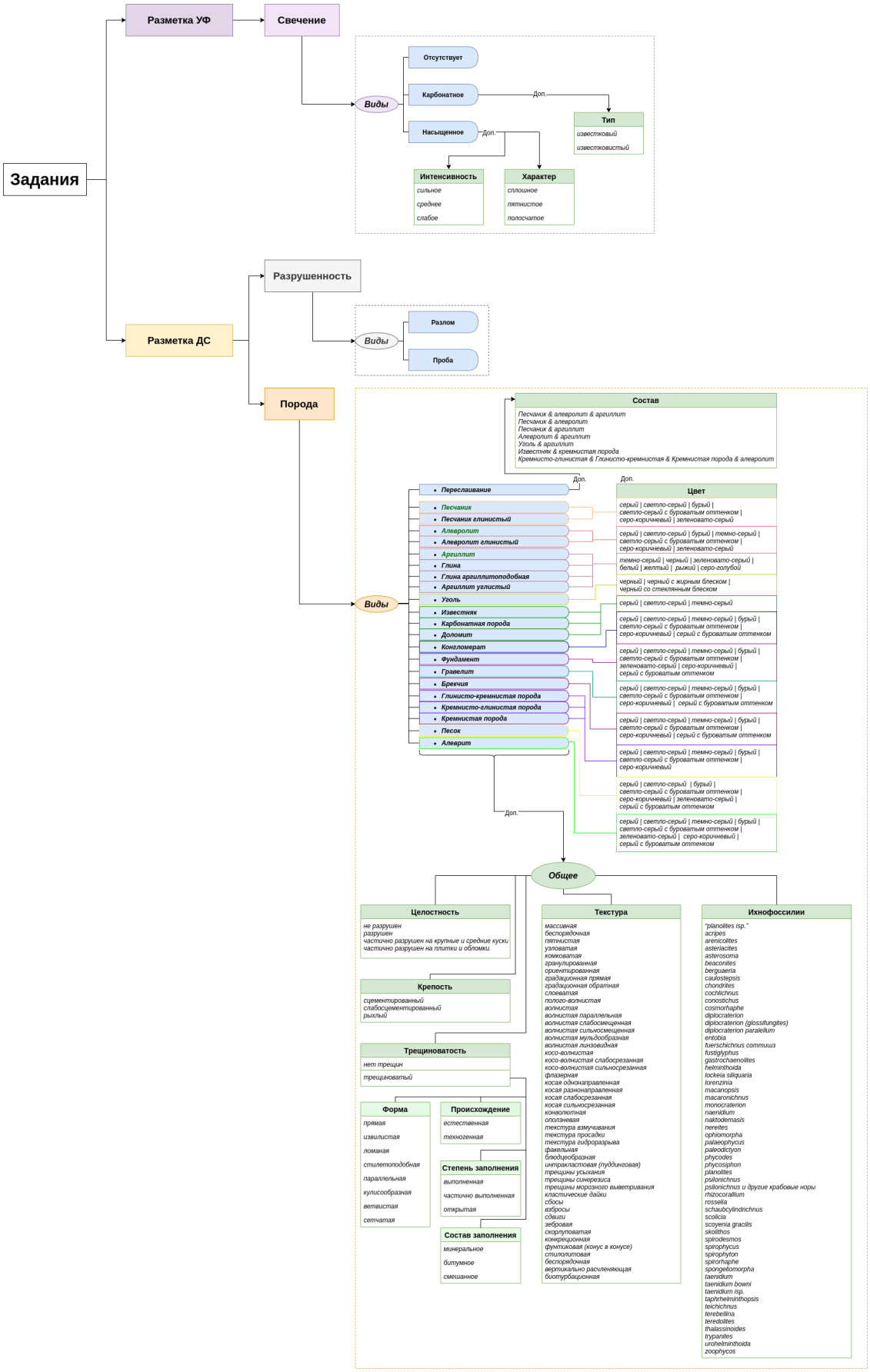


Рис. 15: Схема пользовательского опроса

Глава 3. Проведение исследования

Как уже отмечалось, в данной задаче машинное обучение применяется для решения проблемы классификации. Оценку качества результатов построенных моделей, можно получить с помощью стандартных параметров – точность (precision) и полнота (recall), определяемых как:

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

True positive (TP) – элементы датасета, которые были определены как класс 1, на самом деле являющиеся им в размеченной выборке.
False positive (FP) – элементы датасета, которые были определены как класс 1, на самом деле являющиеся классом 2 (или 3) в размеченной выборке.
False negative (FN) – элементы датасета, которые были определены как класс 2 (или 3), на самом деле являющиеся классом 1 в размеченной выборке.

Класс	Классификация пород	Определение нефтенасыщенности	Определение разрушенности
Класс 1	Песчаник	Нефтенасыщенный	Разрушен
Класс 2	Алевролит	Карбонатный	Не разрушен
Класс 3	Аргиллит	Не насыщенный	

Таблица 1: Классы в подзадачах

В качестве базовой метрики качества модели для каждой подзадачи была выбрана F1-мера, позволяющая одновременно оценить и полноту, и точность результатов предсказания. Поскольку в задаче не было причин в выборе одного из параметра более значимым, оба эти параметра считались одинаково значимыми и поэтому использовалась F1-мера:

$$F1measure = \frac{2 * Recall * Precision}{Recall + Precision}$$

Для всех подзадач итоговые показатели рассчитывались как среднее по

найденным значениям F1-меры соответствующих классов.

В роли базовой модели для сравнения, использовался DummyClassifier – простейший классификатор, учитывающий распределение по классам из обучающей выборки. С его помощью был определен нижний порог для более сложных моделей, а также он использовался для контролирования формирования обучающей выборки.

Для каждой подзадачи выборка разбивалась на обучающую и валидационную в соотношении 8:2. Обучающая выборка использовалась для подбора гиперпараметров, выбора лучшей модели и её тестирования, а на валидационной проверялась лучшая отобранная модель.

Следует заметить, что при обучении моделей использовалась кросс-валидация (перекрёстная проверка), что гарантировало корректную оценку качества каждого алгоритма. Для этого применялся метод StratifiedKFold, позволяющий проверить результаты моделей при разном разбиении обучающего датасета на собственно обучающую и валидационную части. Принцип работы StratifiedKFold продемонстрирован на рисунке 16.

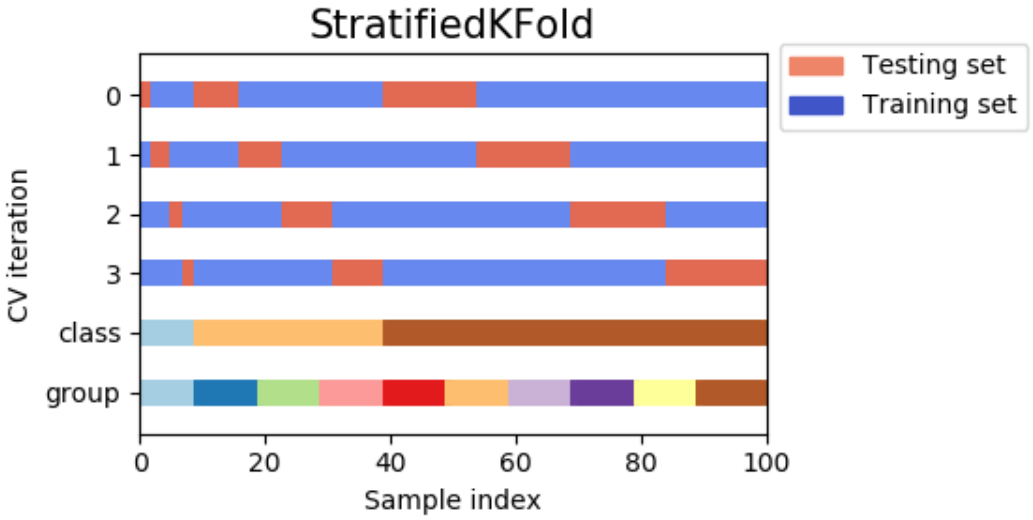


Рис. 16: Принцип разбиения выборки методом StratifiedKFold

Таким образом необходимо было провести исследования по следующим подзадачам:

1. Определение типа породы. Данная задача была сведена к определению трёх самых часто встречаемых пород.

2. Определение карбонатности и нефтенасыщенности. Эти задачи были определены в одну, что связано с корреляцией этих признаков.

3. Определение степени разрушенности.

3.1 Результаты исследования исходных данных

3.1.1 Определение типов пород

Исходя из результатов представленных в таблице 2 для дальнейшего тестирования и улучшения была выбрана модель Extra Tree Classifier. Random Forest Classifier показал аналогичные результаты, однако на обучение требуется больше времени, из-за чего было принято решение в дальнейшем проводить исследования и улучшать модель Extra Tree Classifier. Результаты проверки качества полученной модели на валидационной пред-

Модель	№ проверки	Время на обучение	Время на предсказание	Precision	Recall	F1	Accuracy
Dummy Classifier	1	0.017	0.051	0.32	0.321	0.32	0.321
	2	0.01	0.042	0.32	0.321	0.32	0.321
	3	0.009	0.041	0.32	0.321	0.32	0.321
	4	0.009	0.041	0.32	0.321	0.32	0.321
	5	0.009	0.041	0.32	0.321	0.32	0.321
Random Forest Classifier	1	25.185	2.49	0.759	0.759	0.759	0.759
	2	24.777	1.938	0.759	0.758	0.758	0.759
	3	24.518	2.128	0.78	0.779	0.779	0.779
	4	25.132	2.77	0.77	0.769	0.769	0.769
	5	24.479	2.691	0.757	0.756	0.756	0.757
Extra Tree Classifier	1	8.942	2.453	0.766	0.766	0.766	0.766
	2	9.126	2.119	0.758	0.759	0.758	0.759
	3	9.241	3.904	0.779	0.778	0.778	0.778
	4	9.104	4.009	0.764	0.763	0.763	0.763
	5	9.131	2.117	0.76	0.76	0.76	0.76
SVM Classifier	1	9.725	0.605	0.449	0.347	0.211	0.347
	2	9.791	0.624	0.385	0.365	0.299	0.365
	3	9.908	0.69	0.382	0.394	0.315	0.394
	4	9.837	0.654	0.359	0.391	0.317	0.391
	5	9.781	0.665	0.391	0.323	0.313	0.323

Таблица 2: Результаты работы моделей классификации типов пород

ставлены на рисунке 17.

```
accuracy_score: 0.7722368736514026
recall_score: 0.7667748988690585
precision_score: 0.7677097576353823
f1_score: 0.767101851028093
```

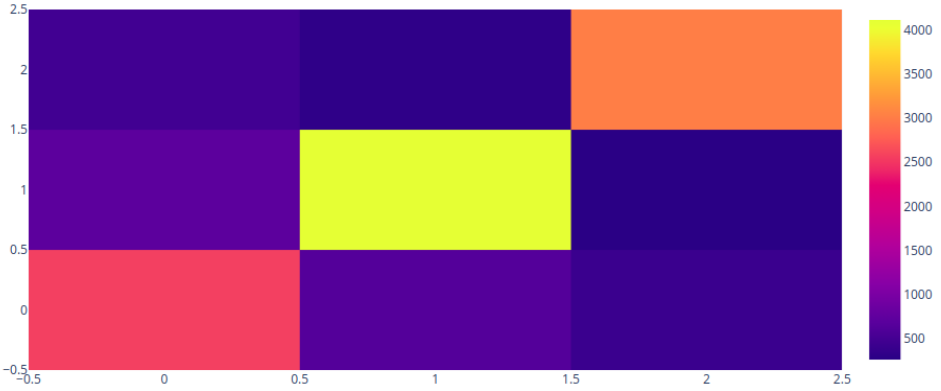


Рис. 17: Матрица ошибок модели на валидационной выборке

3.1.2 Определение типа насыщения

Модель	№ проверки	Время на обучение	Время на предсказание	Precision	Recall	F1	Accuracy
Dummy Classifier	1	0.02	0.041	0.32	0.321	0.32	0.321
	2	0.018	0.042	0.32	0.321	0.32	0.321
	3	0.019	0.041	0.32	0.321	0.32	0.321
	4	0.015	0.044	0.32	0.321	0.32	0.321
	5	0.017	0.051	0.32	0.321	0.32	0.321
Random Forest Classifier	1	27.151	3.391	0.823	0.821	0.822	0.823
	2	25.213	3.43	0.821	0.819	0.82	0.821
	3	26.519	2.918	0.819	0.816	0.818	0.819
	4	23.465	2.783	0.809	0.807	0.808	0.809
	5	26.474	2.891	0.811	0.804	0.807	0.811
Extra Tree Classifier	1	8.942	2.453	0.823	0.821	0.822	0.823
	2	9.126	2.119	0.821	0.819	0.82	0.821
	3	9.241	3.904	0.819	0.816	0.818	0.819
	4	9.104	4.009	0.809	0.807	0.808	0.809
	5	9.131	2.117	0.811	0.804	0.807	0.811
SVM Classifier	1	9.725	0.605	0.449	0.347	0.211	0.347
	2	9.791	0.624	0.385	0.365	0.299	0.365
	3	9.908	0.69	0.382	0.394	0.315	0.394
	4	9.837	0.654	0.359	0.391	0.317	0.391
	5	9.878	0.665	0.391	0.323	0.313	0.323

Таблица 3: Результаты работы моделей классификации типов насыщения

Результаты проверки качества полученной модели на валидационной представлены на рисунке 18.

```
accuracy_score: 0.8288888888888889
recall_score: 0.8288888888888888
precision_score: 0.833951953441614
f1_score: 0.8255294678160208
```

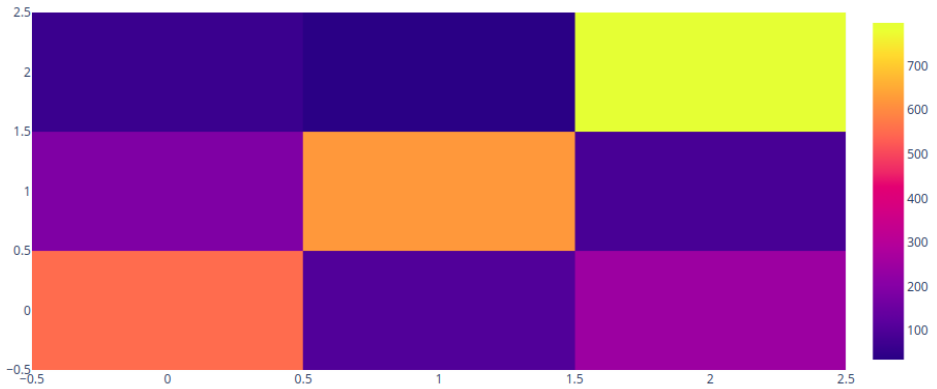


Рис. 18: Матрица ошибок модели на валидационной выборке

3.1.3 Определение трещиноватости

Модель	№ проверки	Время на обучение	Время на предсказание	Precision	Recall	F1	Accuracy
Dummy Classifier	1	0.003	0.034	0.487	0.487	0.487	0.487
	2	0.002	0.025	0.487	0.487	0.487	0.487
	3	0.002	0.02	0.488	0.488	0.488	0.488
	4	0.001	0.019	0.488	0.488	0.488	0.488
	5	0.001	0.02	0.488	0.488	0.488	0.488
Random Forest Classifier	1	3.424	2.594	0.667	0.666	0.667	0.666
	2	3.346	2.192	0.655	0.655	0.655	0.655
	3	3.332	1.782	0.656	0.655	0.655	0.655
	4	3.407	1.648	0.653	0.652	0.652	0.652
	5	3.361	1.864	0.651	0.65	0.65	0.65
Extra Tree Classifier	1	3.818	2.593	0.663	0.662	0.662	0.662
	2	3.379	2.467	0.654	0.654	0.654	0.654
	3	3.302	2.373	0.651	0.65	0.65	0.65
	4	3.326	2.473	0.645	0.645	0.645	0.645
	5	3.297	1.986	0.65	0.649	0.65	0.649
SVM Classifier	1	0.2	0.02	0.302	0.444	0.322	0.444
	2	0.174	0.026	0.655	0.531	0.415	0.531
	3	0.19	0.02	0.69	0.576	0.501	0.576
	4	0.186	0.029	0.36	0.466	0.341	0.466
	5	0.188	0.024	0.33	0.463	0.323	0.463

Таблица 4: Результаты работы моделей классификации трещиноватости

Результаты проверки качества полученной модели на валидационной представлены на рисунке 19.

```
accuracy_score: 0.6890862944162437
recall_score: 0.6890862944162437
precision_score: 0.689116750745759
f1_score: 0.6890737761286094
```

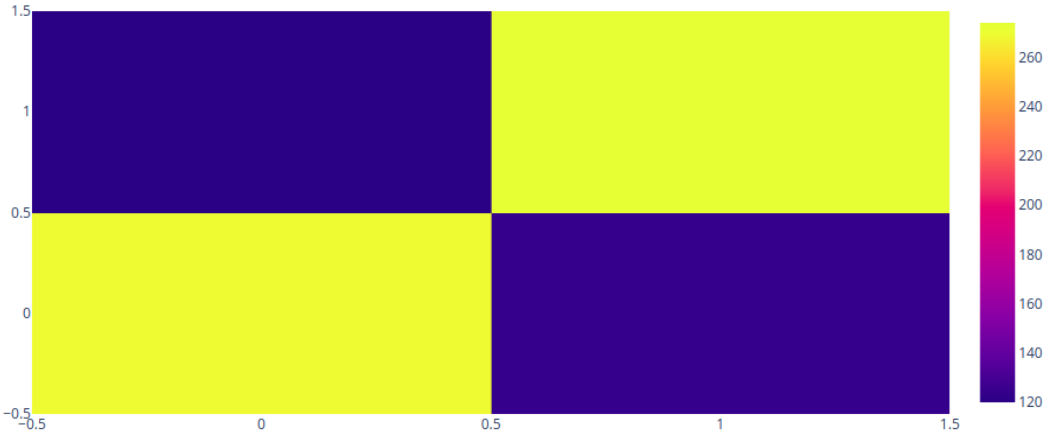


Рис. 19: Матрица ошибок модели на валидационной выборке

3.2 Результаты с использованием обогащенных данных

В результате исследований проведенных на исходной разметке наилучший результат показала модель Extra Tree Classifier. Для проверки гипотез на экспертной разметке было принято решение также использовать модель Extra Tree Classifier, что позволило провести больше экспериментов, благодаря скорости обучения данной модели.

3.2.1 Определение типов пород

Модель	№ проверки	Время на обучение	Время на предсказание	Precision	Recall	F1	Accuracy
Extra Tree Classifier	1	1.121	0.039	0.876	0.87	0.87	0.87
	2	1.319	0.056	0.832	0.832	0.831	0.832
	3	1.185	0.024	0.842	0.841	0.842	0.841
	4	1.133	0.027	0.842	0.84	0.839	0.84
	5	1.146	0.044	0.862	0.859	0.86	0.858

Таблица 5: Результаты работы моделей классификации типов пород

Результаты проверки качества полученной модели на валидационной выборке представлены на рисунке 20 и таблице 6.

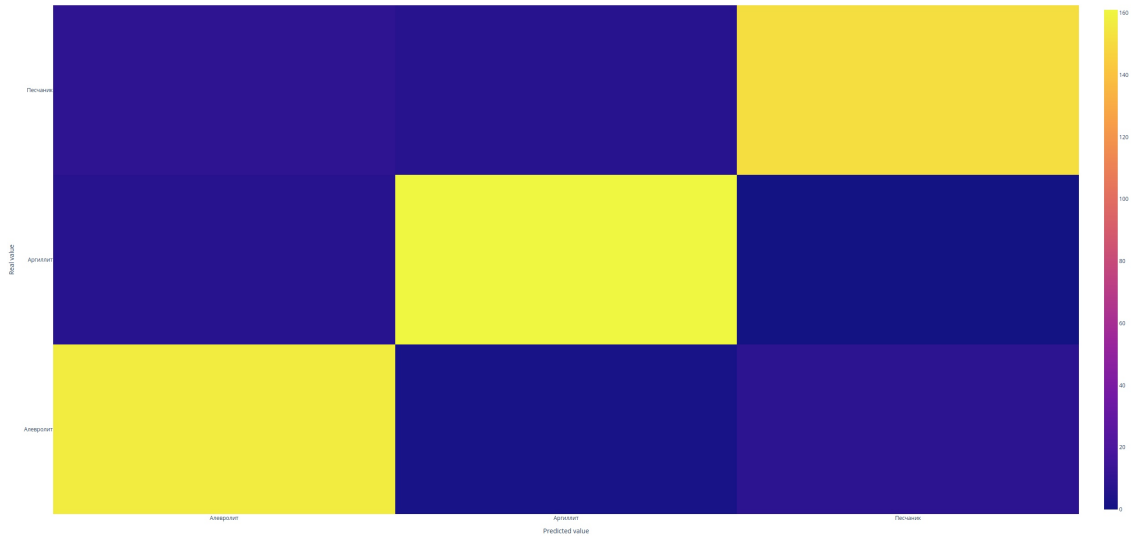


Рис. 20: Матрица ошибок модели на валидационной выборке

Accuracy	Recall	Precision	F1
0.9305555556	0.9305555556	0.9310358914	0.9305076947

Таблица 6: Результаты работы модели на валидационной выборке

3.2.2 Определение типа насыщения

Модель	№ проверки	Время на обучение	Время на предсказание	Precision	Recall	F1	Accuracy
Extra Tree Classifier	1	1.252	0.049	0.875	0.875	0.875	0.876
	2	1.359	0.063	0.899	0.896	0.897	0.896
	3	1.128	0.019	0.926	0.926	0.925	0.925
	4	1.224	0.037	0.883	0.882	0.883	0.882
	5	1.352	0.058	0.889	0.888	0.889	0.888

Таблица 7: Результаты работы моделей классификации типов насыщения

Результаты проверки качества полученной модели на валидационной выборке представлены на рисунке 21 и таблице 8.

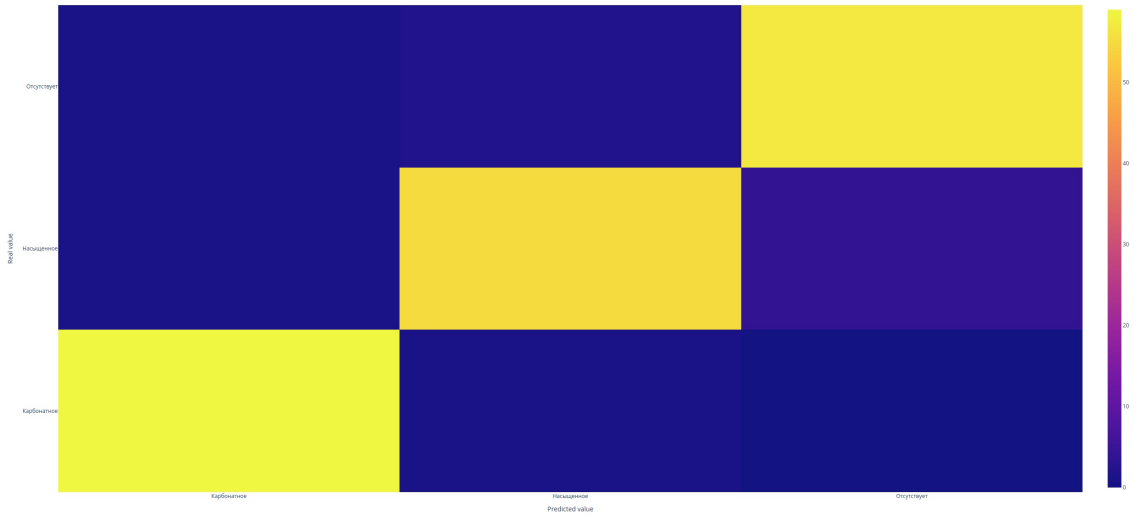


Рис. 21: Матрица ошибок модели на валидационной выборке

Accuracy	Recall	Precision	F1
0.95	0.95	0.9499717354	0.9498529206

Таблица 8: Результаты работы модели на валидационной выборке

3.2.3 Определение трещиноватости

Модель	№ проверки	Время на обучение	Время на предсказание	Precision	Recall	F1	Accuracy
Extra Tree Classifier	1	2.718	1.493	0.753	0.752	0.752	0.752
	2	2.479	1.467	0.744	0.743	0.744	0.744
	3	2.202	1.373	0.741	0.742	0.741	0.741
	4	2.426	1.373	0.735	0.734	0.735	0.735
	5	2.397	1.486	0.745	0.733	0.741	0.739

Таблица 9: Результаты работы моделей классификации трещиноватости

Результаты проверки качества полученной модели на валидационной выборке представлены на рисунке 22 и таблице 10. Не смотря на улучшение метрик при перекрёстной проверке, на валидационной выборке были получены результаты аналогичные результатам на исходной разметке. Основная проблема в определении мест взятия пробы как разрушенных, что эксперты не всегда корректно размечали на исходных фотографиях. Также после анализа полученных результатов был сделан вывод о том, что нет строгой границы в размере разрушенной области, у разных экспертов получались разные результаты при разметке разрушенности, так как некоторые отмечали мельчайшие трещины, а другие ограничивались разметкой только крупных разломов.

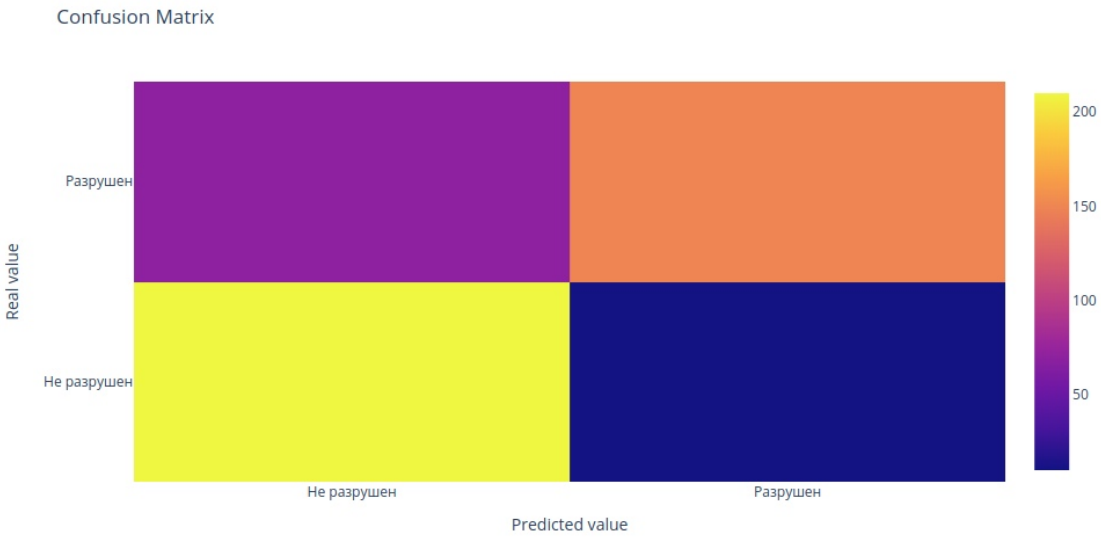


Рис. 22: Матрица ошибок модели на валидационной выборке

Accuracy	Recall	Precision	F1
0.71529	0.71529	0.7299717354	0.728529206

Таблица 10: Результаты работы модели на валидационной выборке

Заключение

В рамках проведённых работ были выполнены поставленные задачи в п.1 и проведены следующие работы:

- Исследована имеющаяся база данных фотографий керна и осуществлена чистка от неподходящих по размеру и некорректно размеченных фотографий.
- Разработан подход к обработке изображений, состоящий из первичной сегментации и последующего перехода к табличному виду с помощью векторов признаков.
- Решены 3 подзадачи классификации: идентификация типа породы, выявление источника УФ свечения и определение разрушенности образцов, проведены соответствующие исследования и продемонстрированы результаты.
- Для каждой подзадачи проведён сравнительный анализ алгоритмов по достижению целевой метрики F1-меры, по результатам которого в каждом случае лучшим был признан Extra Trees Classifier.
- Реализован вспомогательный прототип ПО, позволяющий проводить ручную сегментацию фотографии керна для дальнейшего описания экспертом.
- Собран обогащенный датасет, с строго локализированными областями и с детальным описанием более 12000 фотографий керна.
- Проведен анализ полученной экспертной разметки и показано улучшение результатов при использовании моделей машинного обучения.

Перспективы развития

Полученные результаты использованы для создания веб-сервиса типизации фотографий керна, который был интегрирован во внутреннюю инфраструктуру компании ПАО «Газпром нефть» в качестве компоненты

системы поддержки принятия решений. Также планируется дальнейшее развитие веб-приложения для ручной разметки фотографий керна и его внедрение во внутренние системы и сторонние организации, в том числе в образовательных целях. По разработанной схеме описания слоев керна появились инициативы для сбора дополнительных данных с использованием разработанного сервиса. Помимо разработанных программных продуктов ценность представляет и собранный набор обогащенных данных, который можно использовать для улучшения алгоритма сегментации, обучения нейронных сетей, выделение особенностей экспертной разметки и предиктивной аналитики дополнительных собранных параметров, помимо целевых.

Список литературы

[1] Воробьев К.А. Воробьев А.Е. Тчаро Х. Цифровизация нефтяной промышленности: технология «цифровой» керн // Вестник Евразийской науки, – 2018. – №3.

[2] Алтунин А.Е., Мальшаков А.В., Семухин М.В., Ядрышникова О.А. ООО ТННЦ. Методы компьютерной обработки фотографий керна при изучении коллекторских свойств продуктивных пластов. // Нефтяное хозяйство – 2013. – №11. – с. 12-16.

[3] Чистяков С.П. СЛУЧАЙНЫЕ ЛЕСА: ОБЗОР // Труды Карельского научного центра РАН – 2013. – №1. – с. 117-136.

[4] Geurts, P., Ernst, D., Wehenkel, L. Extremely randomizedtrees // Machine Learning. – 2006. – №63. – p. 3–42.

[5] Srivastava D., L. Bhambhu Data classification using support vector machine. // Journal of Theoretical and Applied Information Technology. – 2010. – №12(1). – p. 1-7.

[6] Н.М. Недоливко. Исследование керна нефтегазовых скважин // Изд-во ТПУ, – 2006. — 170 с.

[7] Center Computer Science. Анализ изображений и видео. Глобальные признаки. 2017— Режим доступа: `https://compscicenter.ru/media/courses/2017-autumn/spb-images-and-video-1/slides/images_and_video_1_lecture_031017.pdf` (дата обращения: 14.12.2019). – Текст: электронный.

[8] Официальный сайт документации Python: сайт. – URL: `https://www.python.org/` (дата обращения: 21.02.2020). – Текст: электронный.