Санкт-Петербургский государственный университет

Математическое обеспечение и администрирование информационных систем

Информационные системы и базы данных

Ерзикова Юлия Ивановна

Применение методов машинного обучения к задаче классификации фаций на основе данных каротажа

Бакалаврская работа

Научный руководитель:
к. ф.-м. н., доцент Графеева Н.Г.

Рецензент:
асс. Волчек Д.Г.

Санкт-Петербург

2019

Saint-Petersburg State University

Software and Administration of Information Systems

Information Systems and Data Bases

Erzikova Iuliia

Application of machine learning methods to the problem of facies classification from well logs

Bachelor's Thesis

Scientific supervisor:
PhD, associate professor Natalia Grafeeva

Reviewer:
Ass. Dmitry Volchek

Saint-Petersburg

2019

Оглавление

Введение

Согласно статистическим данным [44], предоставленным информационно-консалтинговой компанией Enerdata, в настоящее время мировое энергопотребление находится в восходящем тренде (Рисунок 1). Соответственно, возникает потребность оперативной разработки новых месторождений полезных ископаемых, а также более рационального подхода к эксплуатации уже существующих. Высокие производственные показатели добычи природных ресурсов обеспечиваются выполненной своевременно и на должном уровне интерпретацией разведочных данных, а именно, измерений различных свойств горных пород, полученных в ходе геофизических исследований скважин.

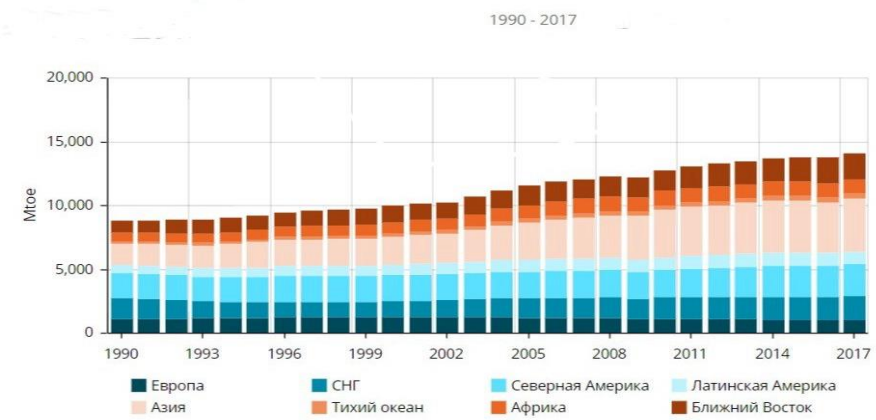

Рисунок 1. Тенденция роста мирового энергопотребления, наблюдаемая за 1990 - 2017 года.

Основополагающим способом изучения и анализа петрофизических свойств (комплекс физических свойств: пористость, плотность, радиоактивность, проницаемость, удельное электрическое сопротивление и т.д.) горных пород является отбор кёрна. Тем не менее, такой подход может быть применим не всегда (неполный вынос кёрна на поверхность) или же он часто не дает полного представления об исследуемых объектах, это обусловлено тем, что при отборе и выносе кёрна на поверхность свойства породы существенно изменяются [40]. Более того, отбор кёрна нерентабелен в связи с высокими экономическими затратами на его применение.

Однако, существует альтернативный метод извлечения важной информации о физических свойствах пород, родоначальниками которого являются братья Шлюмберже. В 20–30-е годы XX века они придумывают технологию, частично или полностью заменяющую отбор кёрна, которая позволяет измерять разнообразный спектр физико-химических свойств пород непосредственно в скважине на разных глубинах с помощью проведения

специальных геофизических исследований. Измерения производятся посредством оборудования, содержащегося внутри геофизического зонда, опускаемого в скважину. Такие исследования для выявления и добычи природных ресурсов известны как каротаж, данный способ разведки является одним из самых популярных на сегодняшний день. Интересным историческим фактом является то, что изначально братья Шлюмберже измеряли лишь электрическое сопротивление под поверхностью земли, для того чтобы искать металлическую руду, но со временем развилась идея измерения целого набора свойств различных ископаемых прямо из скважин.

В настоящее время результаты внутрискважинных исследований, проводимых с целью выявления оптимальной схемы разработки месторождений, базируются на некоторых эмпирических знаниях о предметной области, измерениях, полученных в ходе этих исследований, и физическом моделировании месторождения в результате совокупной интерпретации измерений, которая производится экспертами - петрофизиками. Однако, конструирование таких моделей весьма трудозатратно, требует привлечения большого количества высококвалифицированных специалистов, находящихся сейчас в дефиците. Более того, зависимости числовых характеристик, полученных в результате каротажа и имеющих различную физическую природу, от соответствующих им пород зачастую описываются сложными нелинейными связями, а эксперту достаточно сложно сделать обоснованные предположения относительно изучаемых объектов. Эти умозаключения позволяют выдвинуть предположение о том, что применение методов машинного обучения (Machine Learning) и его отдельной ветви развития - глубокого обучения (Deep Learning), позволяющих найти в исходных данных нетривиальные и потенциально полезные знания, для автоматического анализа и интерпретации каротажных измерений может оказаться весьма эффективным решением.

Рациональность и актуальность вышеобозначенной идеи обусловлена тем, что методы машинного обучения обрели широкую известность за последние несколько лет, поскольку они оказались успешно применимыми для решения задач из всевозможных отраслей. В частности, методы машинного обучения позволяют автоматически оперативно обрабатывать большие объемы данных во многих задачах добывающей промышленности [19, 49, 33], что значительно превосходит возможности человека. Крупные компании, которые занимаются разведкой и разработкой залежей природных ископаемых, активно применяют эти методы, поскольку сейчас основной целью таких компаний является оптимизация технологического процесса формирования месторождений, снижение затрат на него. Так, например, в годовом отчете “Schlumberger” значится, что доход компании за 2017 год увеличился в сравнении с

предыдущим годом и составил $30.440.000, в частности, за счет развития практик применения алгоритмов машинного обучения, стратегий анализа данных, как нового подхода к разработке. Еще одним примером выступает объявленный в 2018 году научно-техническим центром “Газпром нефти” и компанией “Иннопрактика” конкурс “Gazprom neft SmartOil Contest”, целью которого является поиск способов использования технологий машинного обучения для повышения эффективности нефтедобычи. В свою очередь, деятельность международного научного общества геофизиков – исследователей SEG (the society of exploration geophysicists) в последнее время также интенсивно направлена на анализ применимости методов машинного обучения в соответствующей области.

Одной из самых насущных проблем, предстающих перед учеными-геофизиками в ходе интерпретации разведочных данных, является классификация литологических фаций (литофаций), которые предназначены для группировки горных пород со схожими характеристиками. Представление пространственных распределений фаций необходимо для геофизического анализа и 3D-геологического моделирования скважин и месторождений, поскольку насыщение, пористость, проницаемость залежей часто ограничены типами фаций [27,26]. В свою очередь, окончательная геологическая модель играет важнейшую роль в управлении эксплуатацией месторождений и оценке динамики развития добычи полезных ископаемых [25].

В свою очередь, данная важная проблема сводится к одной из фундаментальных задач машинного обучения с учителем, а именно, к задаче мультиклассовой классификации (supervised multiclass classification problem), которая заключается в разбиении наблюдаемых объектов на один из трёх или более заранее известных классов с учётом соответствующих им векторов признаков, а также выявлении зависимости между набором атрибутов (вектором признаков), характеризующим объект, и непосредственно самим объектом. Соответственно, необходима разработка эффективного способа классификации литологических фаций на основе данных геофизической разведки, т.е. каротажа, с помощью методов машинного обучения и, в частности, с помощью глубоких нейронных сетей.

1. Постановка задачи

Целью данного исследования является разработка эффективного способа классификации литологических фаций на основе данных геофизических исследований скважин, т.е. данных каротажа, с помощью методов машинного обучения. Для достижения этой цели в рамках данной работы был сформулирован ряд следующих важных задач:

1. Произвести подробный обзор всех существующих решений за последние 30 лет, посвящённых проблеме автоматической классификации фаций.
2. Найти удовлетворяющий цели исследования набор исходных данных геофизических исследований скважин – данных каротажа. Произвести детальное изучение выбранного набора данных, выявить его ключевые особенности и составить его ёмкое описание.
3. Уделить особое внимание изучению способов предварительной обработки данных геофизических исследований. Осуществить качественную предобработку исходного набора данных с учётом специфики рассматриваемой предметной области - особенностей, выявленных на предыдущем этапе.
4. Выбрать наиболее эффективные методы машинного обучения для решения задачи классификации фаций, опираясь на осуществлённый обзор. Кроме того, определить, какие метрики качества многоклассовой классификации будут использованы в работе для оценки результатов. Для каждого выбранного метода настроить его параметры (выбрать модель) - подобрать оптимальную комбинацию с целью улучшения качества классификации.
5. Провести ряд экспериментов: применить построенные на предыдущем этапе классификационные модели как к обработанным данным каротажа, так и к исходному набору данных без предобработки.
6. Собрать и оценить полученные результаты классификации с помощью выбранных метрик качества. На основе значений метрик сравнить эффективность различных моделей, а также оценить целесообразность произведённой предварительной обработки исходного набора данных каротажных измерений.
7. Проанализировать полученные оценки и сравнения, сделать выводы.

2. Обзор существующих решений

Можно выделить несколько периодов развития подходов к решению задачи классификации фаций на основе геофизических измерений, полученных в процессе каротажных работ.

Работы первого периода (1982-1990 гг.) были основаны на применении классических многомерных статистических методов. Так, в пионерском исследовании [23] 1982 года авторами была разработана новая технология FACIOLOG, в которой используются данные каротажных измерений скважины для ее автоматического зонирования на электрофации на основе применения метода главных компонент (principal component analysis - PCA), позволяющего выбрать из множества всех признаков (измерений свойств пород) те, которые предоставляют информацию для распознавания и характеристики геофизических фаций. Например, включение измерений пластового наклономера и формы звуковой волны в пространство признаков наряду с обычными геофизическими измерениями (Таблица 4) обеспечило выявление значительной структурной информации. С другой стороны, измерения, например, литолого-плотностного каротажа и гамма-спектрометра позволили более точно характеризовать минералогию. В свою очередь, авторы [31] и [18] исследований 1987 года применили дискриминантный анализ (discriminant analysis - DA) для определения литофаций. В научной статье [31] также, как и в [23], уделялось внимание созданию качественного набора данных, включающего измерения, регистрируемые инструментами каротажа (плотность, гамма-излучение, концентрации тория, калия и урана и т.д.), а также заранее идентифицированные на основе петрофизических знаний литологические фации, предназначенные для сопоставления с результатами классификации. Отмечалось, что включение дополнительных каротажных измерений (диэлектрическая проницаемость, фотоэлектрическое поглощение, гамма- спектрометрия и т.д.) повышает качество классификации литофаций. Однако, стоит отметить, что собрать эти измерения не всегда представляется возможным. Основная идея работы заключалась в том, что наблюдение из n измерений на определенном уровне глубины представлялось в виде точки в n-мерном пространстве, а близкие в этом пространстве точки объединялись в кластеры (если концентрация точек была относительно высока), которые в итоге представляли фации. Сопоставление уровней глубины фациальным объектам выполнялось с помощью имеющихся данных и дискриминантной функции (байесовское решающее правило). А в статье [18] был представлен метод идентификации литологий пород с помощью статистического метода дискриминантного анализа данных каротажа скважин месторождения Shublik, в качестве возможного дискриминатора литологий рассматривалась функция литопористости. Подход, описанный в данной работе, был конкретизирован для исследуемого месторождения в виду его характерной особенности - наличия репрезентативных кёрновых данных, которые и использовались для

согласования с каротажными данными, в отличие от [31], где для этой цели использовалась библиотека уже идентифицированных литофаций. Тем не менее, предложенный метод может быть пригоден для исследования любого месторождения, где возможен отбор кёрна. Авторы данной работы оценили несколько построенных на основе 811 наблюдений моделей каротажа и выбрали наиболее подходящую, причем эта модель предназначена для определения лишь трех литологий (известняк, сланец и глинистая порода). Окончательная модель, выбранная для формации Shublik, прогнозировала литологии с точностью в 75%.

Таким образом, в первых работах для решения поставленной задачи применялись классические многомерные статистические методы. Однако, эти методы не являются гибкими и сильно зависят от наличия большого объема статистических данных, что значительно снижает эффективность их применения. Тем не менее, итоги, достигнутые в работах первого периода, послужили мотивацией для проведения дальнейших исследований.

Второй период (1990-2006 гг.) ознаменовался исследованиями [17] 1990 года и [38] 1992 года, авторы которых отошли от рассмотрения многомерных статистических методов, они использовали нейронные сети прямого распространения, обучаемые методом обратного распространения ошибки (feed-forward neural network with backpropagation), для классификации пород. Опытным путем было установлено, что с помощью данного метода зачастую можно эффективнее и значительно оперативнее определять фации, чем это вручную делают эксперты-петрофизики. В свою очередь, сравнительное исследование [32], проведенное с учетом данных каротажа и кёрна из глауконитового сланцеватого месторождения, показало, что результаты классификации фаций, полученные на основе применения нейросетевого подхода, превзошли результаты дискриминантного анализа. Однако также отмечалось, что для проверки справедливости и общности полученного вывода и уточнения выбора модели нейронной сети необходимы дальнейшие исследования в этом направлении. Так, с тех пор нейронные сети были применены ко все большему числу задач интерпретации геофизических разведочных данных. Например, исследователи в своей работе [49], опубликованной в 1996 году, спроектировали модель искусственной нейронной сети с обучением методом обратного распространения ошибки (back-propagation neural network - BPNN) в качестве эффективного способа определения нефтеносных пород и их характеристик, а также выявления точных взаимосвязей между общим содержанием органических веществ в породах и каротажными измерениями. Вместе с тем был применён метод, называемый динамическим созданием узлов (dynamic node creation - DNC), для того чтобы ускорить процесс обучения модели. Впервые модульная искусственная нейронная сеть была применена авторами исследования [2] 2002 года для решения задачи классификации фаций на основе следующих

каротажных измерений: плотность, гамма-излучение, каротаж сопротивления и акустический каротаж, нейтронная пористость. Подход основывался на применении модульной нейронной сети с использованием двухэтапного прогнозирования. Сперва объединенные в ансамбль трехслойные искусственные нейронные сети обратного распространения (BPNN) решали многоклассовую задачу классификации фаций, которая с их помощью эффективно сводилась к ряду двухклассовых классификационных задач. Каждая из таких новообразованных задач решалась соответствующей нейросетью в ансамбле. Затем, с помощью трехслойной рекуррентной нейронной сети эффективно определялись неоднозначные и ложные классификации, осуществленные на предыдущем этапе, и наконец выдавался итоговый результат (один выходной слой). Авторы отметили, что несколько таких модульных нейронных сетей могли быть также скомбинированы в кооператив, тогда итоговое решение о причислении той или иной породы к конкретному классу принималось путем голосования. Эффективность разработанного метода была продемонстрирована сначала с использованием синтетических данных, смоделированных на основе каротажных измерений трех фаций с добавлением случайного шума на уровне 10%, а затем метод был применен для определения четырех различных фаций в пределах месторождения Несс в Северном море. Процент попадания (hit rate) при определении фаций для скважин, не присутствующих в обучающей выборке, в среднем составил около 90 %. В данной работе также было выявлено, что нейросетевой подход, применяемый к поставленной задаче, оказывается эффективнее, чем подход, основанный на дискриминантном анализе.

Итак, для работ второго периода было характерно развитие новых подходов, основанных на использовании различных нейронных сетей. Как отмечают авторы этих исследований, успешное применение нейронных сетей к задачам, связанным с интерпретацией разведочных данных, требует навыков и опыта в определении архитектуры нейронной сети, её настройке. Кроме того, необходимо располагать качественным набором данных достаточно большого объема, а также необходимо уметь справляться с различными проблемами, например, с возможной проблемой переобучения нейросети.

Для третьего периода (2006-2015 гг.) характерно то, что наличие вышеописанных сложностей привело к появлению значимых работ, в которых для решения задачи классификации фаций применяются как подходы на основе нейронных сетей, так и уже иные подходы. Более того, одной из главных целей этих работ стало сравнение эффективности различных применяемых методов. Например, в исследовании [48] 2006 года было показано, что в задаче определения пяти фациальных пород из скважин месторождения “Tensleep Sandstone” на основе данных каротажа (а именно: гамма-излучение, нейтронная пористость, плотность, удельное электрическое сопротивление) наивный байесовский классификатор (naive bayesian classifier - NBC) оказывается более универсальным методом и в целом более точным,

чем решение на основе дискриминантного анализа (DA), примерно на 4%. Также авторы отметили, что NBC не требует предположений о нормальности распределении данных, что делает его более универсальным методом, чем DA. Целью другого исследования [27] 2007 года являлось определение наиболее эффективного классификатора для прогнозирования фаций на основе данных, которые впоследствии стали своего рода эталонными и будут описаны в разделе 3 настоящей работы. Стоит отметить, что авторы [27] являются создателями этого набора данных. Кроме того, ввиду того что каротажные данные постоянно дополняются учеными Канзасского государственного университета, данные за 2007 год существенно отличаются от рассмотренных в разделе 3: они содержат лишь 3600 измерений и 8 выделенных фациальных пород. В [27] были рассмотрены и сравнены между собой семь классификаторов, основанных на четырех различных подходах, а именно, классические параметрические методы (классификаторы линейного и квадратичного дискриминантного анализа, а также классификатор на основе расстояния Махаланобиса) и методы такие как: нечёткая логика, алгоритм k-ближайшего соседа (рассматривалось два различных k-nn классификатора) и нейронная сеть обратного распространения (BPNN). Признаки, использовавшиеся авторами для классификации, представлены в Таблице 4. Несомненным достоинством [27] является то, что для оценки качества классификации авторы использовали разнообразные выразительные метрики, такие как: точность (доля правильных ответов) классификации для всех фаций в целом и отдельно для ключевых трёх фаций (доломит, пакстоун, грэйнстоун), соотношение числа предсказанных ключевых фаций (вместе взятых) к их фактическому числу, стандартное отклонение в восьми соотношениях предсказанных чисел фаций к фактическим для восьми классов, а также точность с учётом смежных фаций (о том, какие фации являются смежными, подробно описано в разделе 3). Итог рассматриваемой работы был таков: BPNN превзошла все другие классификаторы, испытанные в задаче классификации фаций пород, занимая первое место по совокупности значений рассматриваемых метрик. Так, без учета смежных фаций была достигнута точность классификации в 68% для всех фаций, однако для ключевых фаций точность оказалась все же не очень высокой (53%). Два рассматриваемых в работе *k -nn* классификатора были близки друг к другу по достигнутым результатам и заняли второе место по эффективности среди методов. Остальные классификаторы значительно уступают первым трём. Более того, в статье обсуждались возможные причины таких результатов (например, неэффективность статистических параметрических моделей может быть оправдана физической природой данных и наличием в них нелинейных зависимостей), а также предлагались некоторые идеи для их улучшения (например, метод, объединяющий некоторые аспекты непараметрических методов, рассмотренных в статье). Еще одним интересным исследованием стало [1] 2010 года, где нелинейный метод опорных векторов (support vector machine - SVM) применялся для классификации электрофаций. В отличие от

предыдущей работы, данные, задействованные в этом исследовании, значительно меньше по объему и мало изучены (всего 3 скважины и 5 каротажных измерений - гамма-излучение, нейтронная пористость, объёмная плотность, пористость по данным акустического каротажа и кажущееся электрическое удельное сопротивление по индукционному зонду с большим радиусом исследования). SVM-классификатор сравнивался с классификаторами, основанными на линейном дискриминантном анализе (LDA) и вероятностной нейронной сети (probabilistic neural network - PNN). С помощью статистического анализа ошибок, который выполнялся с использованием матрицы несоответствий и ошибок регрессии (коэффициент корреляции, среднеквадратичная ошибка, средняя абсолютная ошибка и максимальная абсолютная погрешность), было установлено, что SVM-подход для классификации имеющихся фаций превзошел подход на основе PNN. LDA оказался неэффективным в данном исследовании. В свою очередь, авторы статьи [6] 2014 года поставили главной целью своего исследования анализ и сравнение пяти подходов на основе искусственных нейронных сетей, которые были применены к задаче идентификации пяти фаций (доломит, известковый доломит, ангидрит, доломитовый известняк и известняк) в четырех скважинах месторождения Южный Парс в Иране с помощью данных каротажа (гамма-излучение, нейтронная пористость, глубинный боковой каротаж, малоглубинный боковой каротаж, фотоэлектрический эффект, плотность и акустический каротаж). В отличие от предыдущих работ, в [6] особое внимание уделялось предобработке имеющихся данных, а именно, были указаны три сценария заполнения пропусков в геофизических данных: исключение из рассмотрения строк с пропусками, интерполяция отсутствующих значений и их прогнозирование специально обученной нейронной сетью. Непосредственно уже для решения задачи рассматривались следующие подходы: нейронные сети, обучаемые методом обратного распространения ошибки (BPNN), нейронные сети радиальных базисных функций (radial basis function - RBF), вероятностные нейронные сети (PNN), нейронные сети конкурентного обучения (competitive learning - CL) и квантизатор вектора обучения (learning vector quantizer - LVQ). Стоит отметить, что в [6] авторы во многом опирались на предыдущие работы [10] и [28], посвященные устройству, видам нейронных сетей и проблеме классификации геофизических фаций соответственно. Результатом работы [6] стал вывод о том, что BPNN оказалась наиболее эффективной в задаче классификации фаций, поскольку генерировала наиболее точные результаты (для используемого набора данных точность классификации 5 фаций составила 74.8%) в сравнении с рассмотренными нейронными сетями другого типа (точность не более 65 %). Однако в данном исследовании не только точность являлась показателями эффективности, но также время обучения и простота архитектуры нейронных сетей. А значит, CL и LVQ оказались более продуктивными методами, чем RBF, ввиду меньшего затраченного на обучение времени при эквивалентных точностях. Также было выяснено, что

пропуски, присутствующие в данных, наиболее точным образом заполнялись с помощью нейросетевого подхода, нежели с использованием интерполяции значений. Наконец было отмечено, что одним из главных факторов снижения точности классификации являлось наличие нежелательных шумов, присущих всем данным геофизической разведки, а потому авторы советуют их предварительно устранять в дальнейших исследованиях.

Таким образом, главной особенностью исследований третьего периода стало применение и сравнение различных методов решения рассматриваемой задачи. Авторы работ данного периода пришли к выводу, что наиболее эффективными оказываются подход на основе метода опорных векторов (SVM) и нейросетевой подход (а именно, BPNN).

Стоит также отметить, что в работах первых трех периодов для решения задачи использовались разные наборы данных каротажа. Поэтому формально судить об объективности и динамике улучшения достигнутых результатов не представляется возможным.

За последний четвёртый период (2016-2019 гг.) появились работы, в которых для поиска и разработки наиболее эффективного способа классификации литологических фаций использовался единый набор данных, предложенный авторами [27] и описанный в разделе 3 текущей работы. Так, в работе [3] 2016 года автор описал процесс решения задачи классификации фаций месторождения Панома с помощью метода машинного обучения, а именно, метода опорных векторов (SVM). В работе было уделено особое внимание тому, как исследовать набор данных, а затем создать, обучить и протестировать модель машинного обучения, в частности, как оптимальным образом подобрать параметры метода (выбор модели) с учетом специфики предметной области. Для оценки результатов классификации фаций из тестовой выборки (данные, не участвовавшие в обучении) использовались весьма показательные метрики – доля правильных ответов (accuracy), точность (precision), полнота (recall) и f- мера (f-measure). Хотя в рабочем процессе [3] использовался достаточно простой, наивный подход, автору все же удалось достигнуть неплохих базовых результатов, так, accuracy в среднем составила 43%, precision - 0.47, recall- 0.46 и f-measure - 0.43. Более того, в 2016 году автор данной работы предложил другим исследователям превзойти достигнутый результат: он объявил соревнование по машинному обучению для определения наиболее точного подхода к идентификации фаций месторождения Панома, которое финансировалось Обществом геофизиков – исследователей (SEG). Тем самым автор [3] задал своего рода тренд на использование в дальнейших исследованиях набора данных, предложенного авторами [27]. По итогам соревнования, описанным его авторами в [4], среди 10 лучших команд из 40 участвующих, для решения задачи использовался

метод машинного обучения, известный как экстремальный градиентный бустинг над решающими деревьями (extreme gradient boosting – XGBoost), который превзошёл подходы, основанные на глубоких нейронных сетях, а точность классификации фаций для тестового набора данных команды-победителя составила в среднем 63.88% (кроме того отмечалось, что случайный лес (Random Forest) также оказался эффективным для решения задачи). Этот результат был достигнут не только за счёт удачного выбора модели, но и за счёт применения техники “создания признаков” (feature engineering), что предполагало использование знаний предметной области для создания дополнительных признаков, которые повышают точность модели классификации. Эти новые признаки основывались на факте о том, что признаки в исходных данных имеют некоторую пространственную корреляцию (по глубине). Однако, авторы исследования [12] 2018 года для тех же исходных данных сумели превзойти лучший результат [4], причем их решение основывалось на использовании нейронной сети глубокого обучения. В данной работе, как и в [4], отмечалось, что каротажные измерения скважин предоставляют только ограниченное количество типов данных, а значит, задача заключается в максимальном извлечении информации (т.е. создании новых признаков), которая предсказывает появление свойств пород, наблюдаемых в кёрне. В большинстве предыдущих исследований [27, 23, 38, 32, 48] используются каротажные данные только на определенной глубине для обучения или тестирования моделей и не учитываются измерения, проведенные на соседних глубинах, или же измерения из разных скважин, но взятых в пределах одной и той же циклической пластовой единицы, а значит, подходы, рассматриваемые в этих работах, имеют существенный недостаток - потерю почти всей пространственной и контекстной информация о фациях. В свою очередь, новый подход к разделению данных, учитывающий важную стратиграфическую информацию о наличии циклического чередования фаций (cyclic facies alternation - CFA), был предложен авторами данного исследования. В дополнение к этому, авторами была выбрана рекуррентная нейронная сеть для решения задачи классификации фаций. Однако было отмечено, что для традиционных рекуррентных нейросетей существует проблема исчезающих градиентов (быстрая потеря информации с течением времени), но особая их разновидность, называемая нейронной сетью с долгой кратковременной памятью (long short-term memory - LSTM), способная к обучению долговременным зависимостям, решает её. Поскольку информация о CFA может распространяться в любом направлении в скважине (сверху вниз или снизу вверх), для проведения исследования была использована двунаправленная нейронная сеть с долгой краткосрочной памятью (bi-directional long short-term memory - BLSTM), которая соединяет два скрытых слоя LSTM противоположных направлений к одному выходному слою. Кроме того, в работе был использован метод прореживания (dropout) для того, чтобы избежать переобучения моделей, а также был применён метод Batch Normalization как эффективный способ ускорения обучения нейронных сетей.

Это также повысило эффективность подхода, применяемого авторами данной работы. В итоге было получено 100 обученных моделей, которые и были использованы для прогнозирования фаций на данных, не участвовавших в обучении нейронных сетей. Доля правильных ответов классификации колебалась от 62.5% до 75.4%, а медианное значение составило 70%, что значительно превзошло лучший результат, достигнутый в [4]. Таким образом, полученные авторами [12] высокие результаты подчеркнули важность включения предметных знаний в модели машинного обучения. Кроме того, было показано, что точность прогнозирования фаций коррелирует с пропорциями фаций в обучающих и тестовых наборах данных. Например, точность классификации наименее распространенных фаций (SS, MS из Таблицы 3) составила 0%. Эти фации наблюдались <1% в тестовом наборе данных и только ~ 5% в тренировочном наборе данных. В то время как определение фаций CSiS, SiSh и D (Таблица 3) с одинаковыми частотами встречаемости в обоих наборах данных (20.40%, 5.87% и 3.06% в обучающем наборе данных; 20.06%, 5.42%, 2.93% в тестовом наборе данных) было осуществлено с точностями 74.05%, 90.00% и 77.78% соответственно. Авторы также отметили, что важно понимать происхождение неправильных определений фаций. Для выявления скрытых тенденций в данных использовался метод главных компонент (PCA), а также продвинутый метод обучения без учителя, т.н. t-распределенное стохастическое соседнее вложение (t-distributed stochastic neighbor embedding - t-SNE), который показал лучшую эффективность, чем линейный PCA. Результаты применения этих методов свидетельствовали о сильном влиянии признака - индикатора NM_M из Таблицы на выделение классов 4 (лишь 3 из 9 фаций являются не морскими, остальные - морскими). Таким образом, был сделан еще один важный вывод о том, что семь из восьми признаков, присутствующих в исходных данных, кроме индикатора, не имеют достаточной информации для совершенно однозначной классификации фаций. Еще одним значимым исследованием для существующей проблемы автоматической интерпретации геофизических данных стала работа [46] 2019 года, главной целью авторов которой являлась разработка эффективной модели, основанной на глубоком обучении, для классификации фаций из скважин месторождения Панома. В этом исследовании рассматривалась новая модель одномерной сверточной нейронной сети (1D-CNN), обучаемой на различных алгоритмах оптимизации: метод адаптивного градиента Adagrad, метод адаптивного шага обучения Adadelta и метод адаптивной инерции Adamax. При этом, как и в [12], авторами был применен метод Batch Normalization, чтобы увеличить скорость обучения 1D-CNN, и dropout, чтобы избежать проблемы переобучения модели. Стоит отметить, что перед началом обучения модели авторы уделили внимание предобработке данных: поскольку значения измерений фотоэффекта не были доступны для некоторых скважин, то его среднее значение по всем скважинам было применено для заполнения пропусков. Все входные данные были преобразованы к значениям со стандартным

нормальным распределением. Для тестирования обученной нейросети использовались данные, не участвовавших в тренировке модели - они составили 20 % от общего набора данных. Сначала авторы [46] сравнили результаты применения к задаче классификации фаций моделей на основе трёх оптимизаторов и пришли к выводу, что модель 1D-CNN, обученная с помощью оптимизатора Adagrad, превосходит модели, обученные на основе Adamax и Adadelta, по точности классификации как для обучающей выборки (96.44%) , так и для тестовой (76.97%) и является более стабильной с увеличением числа итераций обучения. Кроме того, предложенная модель на основе сверточной нейронной сети с использованием трех оптимизаторов сравнивалась с подходом на основе рекуррентной нейронной сети (RNN), с моделью нейронной сети с долгой кратковременной памятью (LSTM), а также с методом опорных векторов (SVM) и методом k-ближайшего соседа (k-nn). 1D-CNN (Adagrad) показала лучшие результаты в сравнении с остальными методами. Эти выводы были получены на основе значений двух метрик качества: accuracy и f-measure. Так, в данном эксперименте accuracy определения фаций без учёта смежных фаций в среднем составила 76.87%, а f-measure - 76.78%. Значения метрик классификации фаций с учётом смежных были таковы: accuracy - 95.54%, f-measure- 95.54%. С высокой accuracy были определены следующие фации: SS, CSiS, D, PS и BS (Таблица 3) - 84.09%, 77.63%, 86.67%, 78.15% и 88.89% соответственно. Кроме того, были показаны различия между рассмотренными моделями: на основе t-критерия Стьюдента было выяснено, что предложенная модель 1D-CNN (Adagrad) показала статистически более значимое улучшение классификации фаций, чем RNN, LSTM и k-nn. В свою очередь, для классификации с учётом смежных фаций модель 1D-CNN (Adagrad) оказалась статистически значительно лучше, чем RNN и SVM.

Итак, характерной особенностью исследований четвертого периода стало использование общего (эталонного) набора данных каротажных измерений и единообразных классических метрик качества автоматической классификации (accuracy, f-measure). Эта особенность выгодно отличает 4 период от предыдущих, поскольку она позволяет осуществить объективное сравнение результатов, достигнутых в работах данного периода. Так, результаты классификации на основе эталонного набора данных для ключевых работ 4 периода были обобщены и представлены в Таблице 1 и в Таблице 2. На их основании можно отчетливо наблюдать динамику постепенного улучшения значений метрик качества классификации. Соответственно, самым высоким оказался результат последней работы [46], в которой использовалась модель сверточной нейронной сети глубокого обучения (1D-CNN), обучаемая на различных алгоритмах оптимизации.

Статья	Подход	Метрики	
		Accuracy	***F-measure***
B. Hall [3] (2016)	SVM	43%	43%
B. Hall, M. Hall [4] (2017)	XGBoost	63.88%	—
J. Jiajun, C.J. Scott, C.A. Stacy [12] (2018)	BLSTM NN	70%	—
Y. Imamverdiyev, L. Sukhostat [46] (2019)	1D-CNN	76.87%	76.78%

Таблица 1. Сравнение результатов классификации.

Статья	Подход	Метрики	
		Accuracy	***F-measure***
B. Hall [3] (2016)	SVM	—	88%
Y. Imamverdiyev, L. Sukhostat [46] (2019)	1D-CNN	95.54%	95.54%

Таблица 2. Сравнение результатов классификации с учётом смежных фаций.

Примечание: результаты работы [12] на момент с 01.05.2019 являются не действительными (статья для публикации была изъята) – это было установлено в ходе общения с авторами данного исследования (причина заключается в том, что авторы неправильно подошли к работе с имеющимися данными, реальное значение *accuracy* составляет лишь 58 %). По этой причине, на данный момент работу [12] можно считать не актуальной.

3. Описание исходных данных

Исторически сложилось, что, начиная с некоторого времени (примерно с 2016 года), в большинстве научных работ, посвященных решению задачи автоматической классификации литологических фаций на основе данных каротажных измерений, для осуществления исследований использовался набор данных, находящийся в открытом доступе на сайте Канзасского государственного исследовательского университета (http://www.people.ku.edu/~gbohling/EECS833/). Поскольку именно на вышеобозначенных данных проводились многочисленные эксперименты, собирались оценки качества полученных результатов классификации, а также сравнивались достигнутые в разных работах результаты, в текущем исследовании именно эти данные будут описаны с указанием присущих им ключевых характеристик и особенностей и использованы для проведения экспериментов.

Областью исследования являются газовые месторождения Хьюготон и Панома, расположенные на юго-западе Канзаса и северо-западе Оклахомы вдоль нефтегазоносного бассейна Анадарко (Рисунок 2). Эти месторождения составляют самую обширную зону в Северной Америке, на которой добывается газ (добыча 963 млрд. $м^3$ газа из 412 000 скважин). На данный момент наиболее приоритетной задачей является геофизическое моделирование месторождения Панома. Для этого, в первую очередь, необходимо найти эффективный способ автоматической классификации фаций на основе измерений, полученных в процессе каротажа скважин этого месторождения [27, 26].

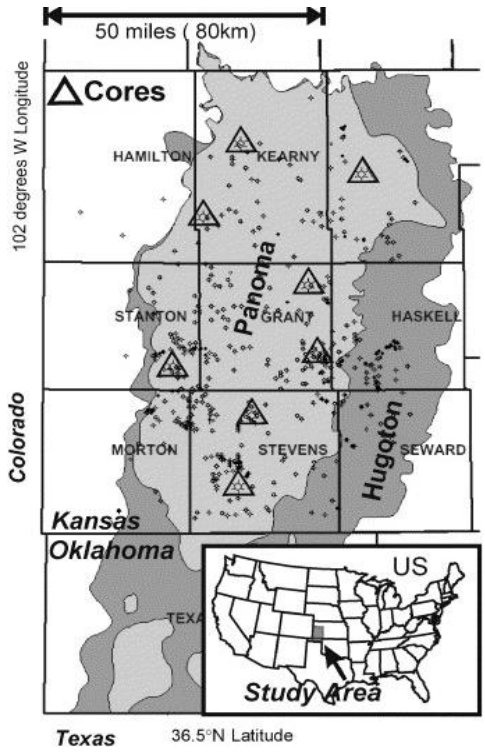

Рисунок 2. Карта, иллюстрирующая контуры исследуемых месторождений [6]. Отмечено около 515 скважин.

В данных содержится информация о том, что исследования проводились в 10 скважинах (SHRIMPLIN, ALEXANDER D, SHANKLE, LUKE G U, KIMZEY A, CROSS H CATTLE, NOLAN, RECRUIT F9, NEWBY и CHURCHMAN BIBLE) месторождения Панома. Всего было произведено 4149 измерений (т.е. 4149 векторов измеренных признаков), взятых с интервалом по глубине в 0.5 фута (0.15 метра). Каждый вектор был также помечен фациальным типом (классом) на основе доступных кёрновых данных из скважин. Осадочные породы в этом исследовании подразделяются на 9 фаций (классов), которые представлены в Таблице 3. Стоит отметить, что псевдо - скважина “RECRUIT F9” специально была введена для улучшения точности классификации девятой фации BS (“филлоидно - водорослевый дефлектор”) [27]. Это было необходимо, поскольку другие скважины содержат лишь малое число фаций BS, что затрудняет их определение при классификации. Согласно [27,26], указанное количество различных фаций является минимальным, необходимым для точного представления физической изменчивости месторождений и классового различия пород по основным петрофизическим свойствам. Фации были выделены на основе породного типа и текстуры. Породы месторождения Панома состоят из различных количеств четырех минеральных компонентов - кальцита, доломита, кварца и глины, относительная пропорция этих минералов и определяет основной тип породы. Кроме того, около 80 % представленных фаций имеют морское происхождение, в свою очередь оставшиеся 20 % имеют континентальное происхождение.

Метка фации	Тип фации	Смежные фации
SS	Nonmarine sandstone (Не морской песчаник)	CSiS
CSiS	Nonmarine coarse siltstone (Не морской крупный алевролит)	SS, FSiS
FSiS	Nonmarine fine siltstone (Не морской мелкий алевролит)	CSiS

SiSh	Marine siltstone and shale (Морские алевролит и сланец)	MS
MS	Mudstone (Аргиллит)	SiSh, WS
WS	Wackestone (Ваккит)	MS, D, PS
D	Dolomite (Доломит)	WS, PS
PS	Packstone-grainstone (Пакстоун-грэйнстоун)	WS, D, BS
BS	Phylloid-algal bafflestone (филлоидно-водорослевый дефлектор)	D, PS

Таблица 3. Справочник фаций с указанием смежных.

Особенностью этого набора данных служит тот факт, что некоторые фации являются смежными, поскольку они близки друг к другу с точки зрения схожести их петрофизических свойств. Это является основной причиной постепенного смешивания таких фаций, а значит, отнести их к правильному классу при классификации довольно сложно (велика вероятность ошибочной классификации). Ввиду этого можно считать, что наличие классификации фаций, близкой к фактической (с учётом смежных фаций – т.н. классификация смежных фаций), является удовлетворительным.
Например, если какой-то объект с соответствующим вектором признаков и фациальной меткой CSiS будет отнесён к классу SS или классу FSiS, то такая классификация не будет считаться ошибочной. В третьем столбце Таблицы 3 для каждой фации указаны смежные с ней.

Горные породы обладают многочисленными физическими и химическими свойствами, которые могут быть использованы для классификации, однако наиболее доступными свойствами для пород, встречающихся в нефтяных и газовых скважинах, являются те, которые измеряются специальными петрофизическими каротажными инструментами, опущенными в скважину после ее пробуривания. Краткое описание измерений свойств фаций, которые присутствуют в рассматриваемом наборе данных и которые могут быть использованы в качестве признаков для классификации методами машинного

обучения, приведено в Таблице 4. Стоит отметить, что в исходных данных представлены 5 основных признаков, соответствующих геофизическим свойствам горных пород - GR, ILD_log10, DeltaPHI, PHIND и PE. Эти характеристики обычно регистрируются для разбуриваемых скважин, начиная с 1970-х годов, и часто используются экспертами для классификации фаций. В свою очередь, 2 важных дополнительных признака (NM_M, RELPOS), полученных из геологической информации, были включены в вектор признаков, характеризующий фацию. Потенциальная значимость этих двух признаков для задачи классификации обусловлена тем, что некоторые фации очень точно определяются значениями данных измерений [27, 26].

Признак	Описание признака	Единицы измерения (для количественных и бинарных признаков)
GR (Gamma ray)	Гамма-излучение - измерение естественной радиоактивности пород в скважине.	API (единица скорости счёта при гамма-каротаже Американского института нефти) GR(API) = GR (мкр/ч) *10.0
ILD_log10 (Resistivity)	Каротаж сопротивления - измерение удельного электрического сопротивления пород (способность препятствовать течению тока).	Омметр (Ωm)
DeltaPHI (Neutron - density porosity difference)	Нейтронный гамма-каротаж: Разность пористости нейтронной плотности. Измерение, коррелирующее с плотностью фации.	%
PHIND (Average neutron-density porosity)	Нейтронный гамма-каротаж: Средняя пористость нейтронной плотности. Измерение, коррелирующее с плотностью фации.	%
PE (Photoelectric effect)	Гамма-гамма-каротаж: Фотоэффект (Фотоэлектрическое поглощение). Измерение излучения электронов фаций, освещенных световыми лучами.	эВ (электронвольт)
Depth	Глубина, на которой производились измерения признаков.	Фут (1 фут = 0.3048 м)
NM_M (Nonmarine / marine indicator)	Показатель того, к какому классу относится фация - морских или не морских.	“1” - не морская, “2” - морская

RELPOS (Relative position)	Относительное положение. Индекс, соответствующий глубине, на которой было проведено измерение (Индекс уменьшается с увеличением глубины, начинается с 1).	—

Таблица 4. Описание признаков, содержащихся в исходных данных.

Комплекс измерений, полученный в ходе геофизических исследований скважин, принято изображать в виде каротажных диаграмм. На рисунке 3 расположен пример такой диаграммы для одной из скважин месторождения Панома, которая была построена на основе измерений физических признаков фаций, имеющихся в рассматриваемом наборе данных (Таблица 4). Исследователи могут использовать такие каротажные диаграммы в качестве очень наглядного способа для интерпретации и оценки результатов классификации, если наряду с имеющимися столбцами изобразить столбец с предсказанными классами. Однако, кроме того, первые 5 столбцов каротажной диаграммы на рисунке 3 наглядно отражают и подтверждают следующий факт: зачастую результаты геофизических исследований скважин отличаются большими уровнями естественных шумов и статистических флюктуаций измеряемых величин [45]. Это объясняется их физической природой, неоднородностями геологических и природных сред, а также неточностями геофизического оборудования для регистрации данных. Зашумленные данные содержат в себе некоторую долю дезинформации (случайные сигналы в виде помех), наличие которой, в свою очередь, негативно отражается на результатах их дальнейшего изучения и интерпретации. Более того, некоторые методы машинного обучения очень чувствительны к наличию шумов в данных (например, метод опорных векторов). Поэтому очистка от шумов и сглаживание данных геофизической разведки является первоочерёдной задачей их обработки и подготовки к интерпретации.

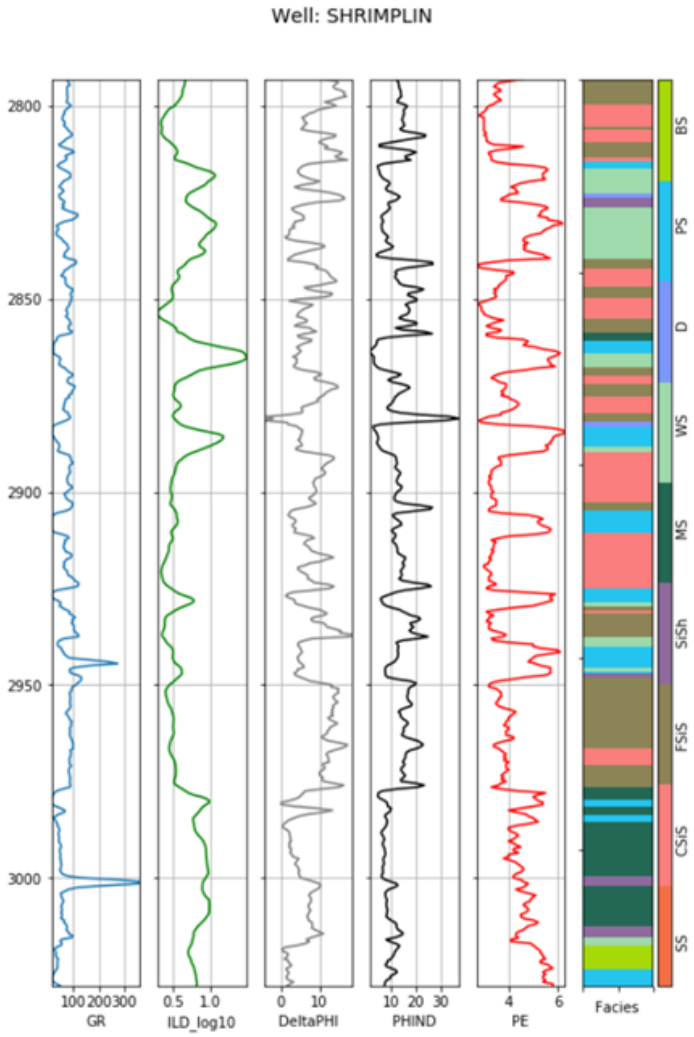

Рисунок 3. Каротажная диаграмма для скважины “SHRIMPLIN” месторождения Панома, отражающая поведение сигналов измерений физических свойств (GR, ILD_log10, DeltaPHI, PHIND, PE) фаций вдоль ствола скважины (т.е. в глубину). Каждому набору измерений на определенном уровне глубины соответствует одна из 9 фаций (самая правая колонка Facies).

Ещё одной важной особенностью, присущей всем данным геофизических исследований скважин, является то, что различные в физическом смысле показатели (признаки) часто различаются по абсолютным величинам (Рисунок 4). Например, в представленном наборе данных каротажных измерений гамма-излучение измеряется в API (единица скорости счета при гамма-каротаже Американского института нефти), каротаж сопротивления - в Ωm (омметр), а фотоэффект измеряется в эВ (электронвольт). В свою очередь, многие методы машинного обучения чувствительны к масштабированию, стандартизации данных (например, метод опорных векторов, метод k-ближайшего соседа, нейронные сети, логистическая регрессия). Ввиду этого, использование таких разнородных признаков в моделях машинного обучения для решения поставленной задачи некорректно без некоторого преобразования, которое бы нивелировало разность их величин.

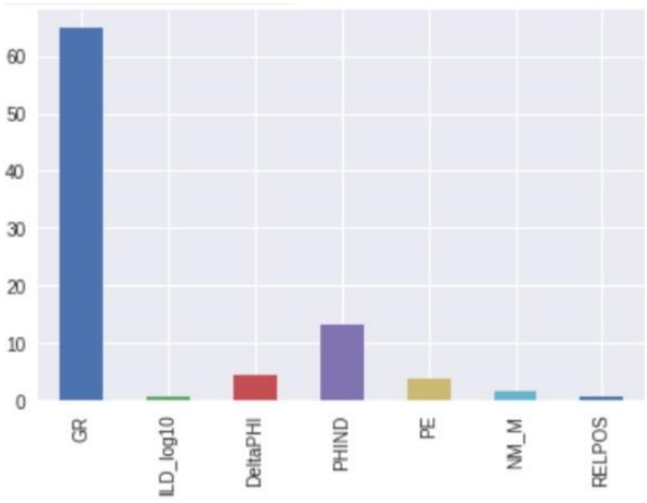

Рисунок 4. Средние значения измерений признаков фаций.

На рисунке 5 представлен небольшой фрагмент описываемого набора исходных данных. При более детальном изучении обнаруживается, что в нём присутствуют пропуски. Так, для трёх из десяти скважин (ALEXANDER D, KIMZEY A и RECRUIT F9) отсутствуют измерения значений фотоэффекта (PE). Поскольку фотоэффект является весомым признаком для осуществления классификации фаций, он всегда измеряется там, где доступен (пропущенные значения объясняются ограниченными измерительными возможностями геофизического оборудования, погружаемого в скважину). Для эффективного решения поставленной задачи классификации методами машинного обучения возникает необходимость заполнения имеющихся в данных пропусков. Тем не менее, эти пропуски составляют около 22% от общего числа измерений (всего 917 отсутствующих значений, при этом более 400 из них идут подряд), а значит, их заполнение соседними значениями невозможно. Кроме того, восстановление пропущенных значений с помощью среднего значения фотоэффекта по всем скважинам нельзя считать удовлетворительным решением, поскольку величина любого имеющегося признака зависит от скважины и глубины, на которой были произведены измерения. На основании этого можно заключить, что в рамках данного исследования возникает дополнительная важная задача корректного восстановления пропущенных измерений фотоэффекта методами машинного обучения.

Facies	Formation	Well Name	Depth	GR	ILD_log10	DeltaPHI	PHIND	PE	NM_M	RELPOS
3	A1 SH	SHRIMPLIN	2795.5	73.97	0.636	14.0	13.385	3.6	1	0.894
3	A1 SH	SHRIMPLIN	2796.0	73.72	0.63	15.6	13.93	3.7	1	0.872
3	A1 SH	SHRIMPLIN	2796.5	75.65	0.625	16.5	13.92	3.5	1	0.83
3	A1 SH	SHRIMPLIN	2797.0	73.79	0.624	16.2	13.98	3.4	1	0.809
3	A1 SH	SHRIMPLIN	2797.5	76.89	0.615	16.9	14.22	3.5	1	0.787
3	A1 SH	SHRIMPLIN	2798.0	76.11	0.6	14.8	13.375	3.6	1	0.7659999
3	A1 SH	SHRIMPLIN	2798.5	74.95	0.583	13.3	12.69	3.7	1	0.745
3	A1 SH	SHRIMPLIN	2799.0	71.87	0.561	11.3	12.475	3.5	1	0.723
3	A1 SH	SHRIMPLIN	2799.5	83.42	0.537	13.3	14.93	3.4	1	0.7020000
2	A1 SH	SHRIMPLIN	2800.0	90.1	0.519	14.3	16.555	3.2	1	0.6809999
2	A1 SH	SHRIMPLIN	2800.5	78.15	0.467	11.8	15.96	3.1	1	0.638

Рисунок 5. Фрагмент исходных данных.

4. Предобработка исходных данных

Качество создаваемых моделей машинного обучения с учителем зависит от многих факторов, в первую очередь – от качества исходных данных. Тем не менее, на сегодняшний день, ввиду непрерывного роста информации, а также разнообразия, сложности и специфичности существующих реальных задач, данные, используемые для их решения, имеют тенденцию быть разнородными, неполными, шумными, непоследовательными. Такие данные искажены и ненадёжны, а значит, не пригодны для создания и обучения моделей машинного обучения [35, 13]. Вот почему предварительная обработка, включающая в себя очистку, нормализацию и преобразование набора данных, а также восстановление пропущенных значений, выявление и отбор признаков для наилучшей характеристики исследуемых объектов, является очень важным этапом в процессе интеллектуального анализа исходных данных – он обеспечивает их дальнейшее эффективное использование в задачах машинного обучения.

Как было выявлено и обозначено в разделе 3 настоящей работы, выбранные для проведения исследования данные каротажа ввиду специфики рассматриваемой предметной области имеют три значимые серьёзные «несовершенства». Следовательно, для более эффективного решения поставленной задачи классификации фаций необходимо осуществить качественную предварительную обработку исходного набора данных, состоящую из следующих этапов:

1) Устранение шумов в каротажных измерениях (для физических признаков GR, ILD_log10, DeltaPHI, PHIND, PE из Таблицы 4).

2) Преобразование данных для устранения сильного различия значений признаков по абсолютным величинам, а также для придания им более нормального распределения.

3) Восстановление неизмеренных значений фотоэффекта для трёх скважин с помощью методов машинного обучения.

Стоит отметить, что в ходе проведённого ранее обзора существующих решений поставленной задачи (раздел 2), было установлено, что авторы разнообразных работ не занимались тщательной предобработкой исходных данных. В данном же исследовании, предобработке уделяется особое внимание, поскольку было выдвинуто предположение о том, что она способна в значительной степени повлиять на качество классификации фаций (в положительном ключе). Это предположение будет проверено в дальнейшем на стадии проведения экспериментов.

4.1. Устранение шумов в каротажных измерениях

Как отмечалось ранее, особенностью измерительных сигналов, регистрируемых с помощью датчиков геофизического зонда при проведении каротажных работ, является то, что они, как правило, нестационарные и значительно зашумлены. Шум — это беспорядочные колебания различной физической природы, отличающиеся сложностью временной и спектральной структуры. Наличие шумов в данных оказывает негативное влияние на их дальнейшую интерпретацию и обработку, поэтому необходимо предварительно подавлять эти шумы.

Применение классических способов очистки сигналов от шумов с помощью различных фильтров (например, фильтры Кальмана, Колмогорова-Винера, сглаживающих, медианных и др.) приводит к искажению формы сигнала (если свойства сигнала априори неизвестны), а это, в свою очередь, может повлечь ошибочную физическую интерпретацию изучаемого процесса [52]. Причём большинство существующих методов анализа зашумленных данных предназначены для линейных и стационарных сигналов. Однако, для изучаемых каротажных измерительных сигналов характерна нелинейность и нестационарность (т.е. естественное изменение их характеристик в пространстве). За последние десятилетия начали развиваться новые методы анализа (например, основанные на вейвлет-преобразовании или преобразовании Фурье, предназначенные для анализа нестационарных, но линейных процессов), однако, в них не учитывается возможность адаптивного формирования базисных функций преобразований, функционально зависимых именно от содержания самих входных данных [52].

В свою очередь, такой подход реализуется в методе эмпирической модовой декомпозиции (empirical mode decomposition - EMD), предложенном Норденом Хуангом в 1995 году и описанном в [30]. Метод является важнейшей составляющей преобразования Гильберта – Хуанга (Huang-Hilbert transform - HHT), получившего широкое применение в некоторых сферах – геофизика, биомедицина. Первоначально HHT и, в частности, EMD использовались при изучении поверхностных волн тайфунов, однако со временем также стали успешно применятся для подавления шумов в данных геофизической разведки [51, 45]. На основании этого, было принято решение использовать EMD для подавления шумов в имеющихся исходных данных каротажа.

Суть метода заключается в том, что анализируемый многокомпонентный (т.е. состоящий из множества колебательных процессов) сигнал $x(t)$ адаптивно раскладывается в сумму L составляющих с различными частотами (первая

составляющая - самая высокочастотная) – эмпирических мод (intrinsic mode functions – IMF) $h^{(i)}(t)$, и остатка $d(t)$: $x(t) = \sum_{i=1}^{L} h^{(i)}(t) + d(t)$.

Причем эмпирические моды не задаются аналитически, они выявляются с помощью итеративного процесса (т.н. отсеивания) и определяются исключительно самим входным сигналом, и обладают следующими свойствами: 1) число максимумов и минимумов функции, а также количество пересечений нуля, должны быть равными или отличаться не более, чем на единицу; 2) среднее значение огибающих, построенных по локальным максимумам и локальным минимумам, близко к нулю. После разложения входного сигнала на моды, необходимо исключить первые, наиболее высокочастотные, составляющие (примерно от 1 до 3), которые обычно и представляют собой шум, содержащийся в сигнале. Основным достоинством EMD в сравнении с классическими методами подавления шумов является то, что исходная форма сигнала искажается в значительно меньшей степени [52].

Метод эмпирической модовой декомпозиции был применён к 5 имеющимся каротажным измерениям (GR, ILD_log10, DeltaPHI, PHIND, PE из Таблицы 4) для каждой из 10 скважин в отдельности. На рисунке 6 представлено сравнение исходного сигнала гамма-излучения с преобразованным сигналом, полученным в результате применения метода эмпирической модовой декомпозиции. В свою очередь, на рисунке 7 представлены две каротажные диаграммы, отражающие общую динамику поведения имеющихся физических измерительных сигналов до и после преобразования. Можно отметить некоторое сглаживание всех 5 сигналов после применения EMD.

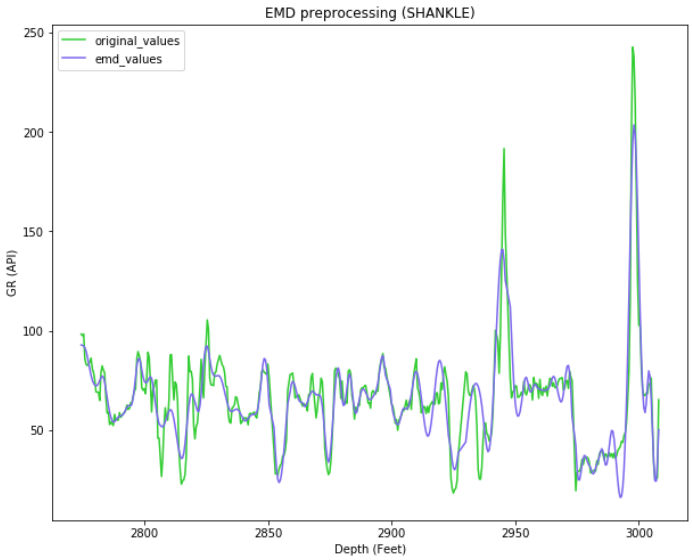

Рисунок 6. Результат применения EMD к сигналу гамма-излучения для скважины SHANKLE.

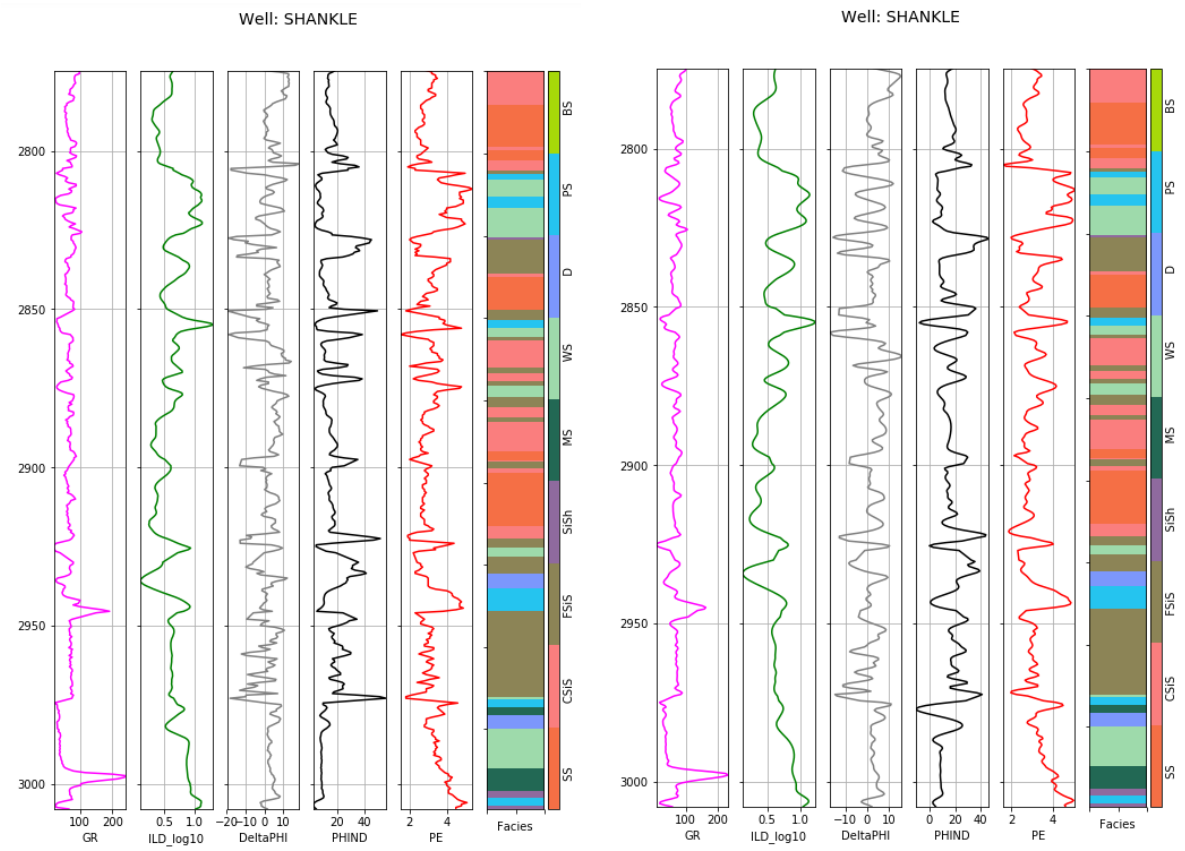

Рисунок 7. Каротажная диаграмма для скважины SHANKLE до применения EMD (слева) и после (справа).

Однако, локальная природа метода EMD может вызывать колебания с очень разнородными масштабами в одной моде или колебаниями с аналогичными масштабами, но в разных модах. Это явление называется смешиванием мод и является нежелательным, поскольку оно всё же приводит к некоторому искажению действительного сигнала. Чтобы преодолеть смешивание мод, был предложен новый метод, называемый ансамблевой (многократной) эмпирической модовой декомпозицией (ensemble empirical mode decomposition – EEMD), описанный в [24, 50]. Новый подход к заключается в многократном добавлении к исходному сигналу в процессе его разложения на моды белого Гауссовского шума и вычислении среднего значения эмпирических мод как конечного истинного результата.

Результат применения метода EEMD к исходным данным каротажа одной из скважин месторождения Панома представлен на рисунке 8. Можно отметить, что новые преобразованные сигналы частично содержат в себе добавленные в процессе декомпозиции белые шумы. Это явление неизбежно (особенно при относительно небольшом количестве итераций в разложении) и считается основным недостатком метода EEMD [14].

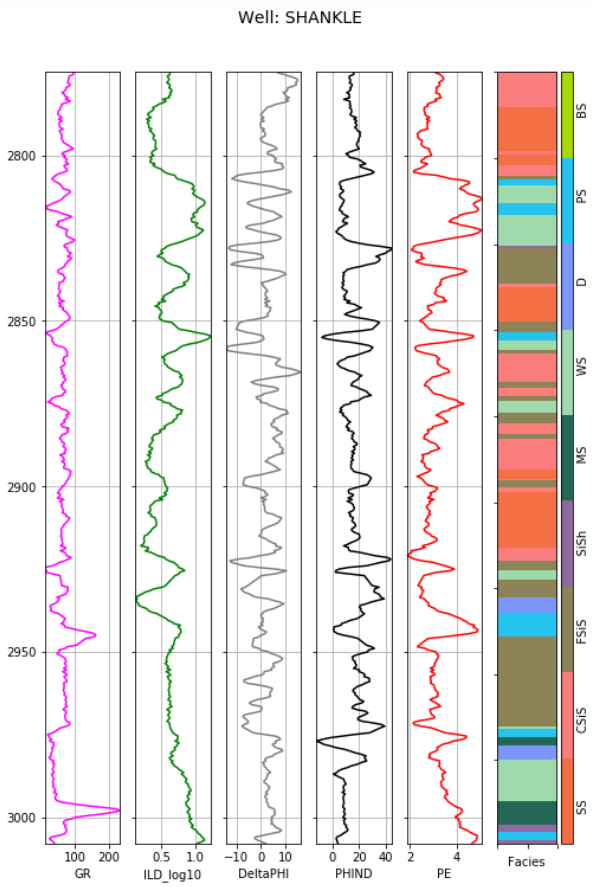

Рисунок 8. Каротажная диаграмма для скважины SHANKLE после применения EEMD.

Для оценки и сравнения эффективности процессов шумоподавления в исходных данных каротажа на основе применения методов EMD и EEMD соответственно была использована среднеквадратичная ошибка (mean squared error - MSE): $MSE = \frac{1}{N}\sum_{i=1}^{N}(x_i - x_i)^2$, где $\tilde{x}-$ преобразованный (очищенный) сигнал, $x-$ исходный сигнал . Небольшое значение метрики MSE является индикатором того, что сигнал $\tilde{x}$ близок к x. Следовательно, чем меньше значение MSE, тем эффективнее можно считать процесс устранения шума (поскольку важно не искажать в значительной степени исходную форму сигнала) [36].

Результаты сравнения представлены в таблице 5, на основании которых можно заключить, что для устранения шумов в исследуемом наборе данных наиболее целесообразным и эффективным является применение классического метода эмпирической модовой декомпозиции (EMD), поскольку он вносит погрешность в каждый из получаемых сигналов в значительно меньшей степени, чем метод многократной эмпирической модовой декомпозиции (EEMD).

Метод / Сигнал	EMD	EEMD
GR	45.5	60.9
ILD_log10	0.002	0.003
DeltaPHI	8.76	10.14
PHIND	22.3	23.7
PE	0.074	0.085

Таблица 5. Значения метрики MSE методов EMD и EEMD, рассчитанные для каждого из 5 имеющихся сигналов (скважина SHANKLE).

4.2. Преобразование данных

После устранения шумов данные необходимо подготовить для решения задач восстановления пропусков и классификации фаций методами машинного обучения, поскольку, как говорилось ранее, они зачастую чувствительны к масштабированию, а признаки, содержащиеся в исходных данных, имеют разную физическую природу. В рамках данного исследования были рассмотрены несколько следующих способов преобразования данных для их дальнейшего использования в различных моделях машинного обучения:

1. Стандартизация (Standard Scaling)

Для каждого признака x каждое его значение на i-ом объекте преобразуется по формуле:

$x_i^{new} = \frac{x_i - \mu}{\sigma}$, где $\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$ – среднее значение признака x, $\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$ – стандартное отклонение признака x, N – длина выборки. При этом предполагается, что x имеет нормальное распределение.

После такого преобразования каждый признак имеет среднее значение, равное 0, и дисперсию, равную 1 (т.е. осуществляется переход от нормально распределённой величины к величине со стандартным нормальным распределением).

2. Линейная нормализация (Min-Max Scaling)

Для каждого признака x его новое значение на конкретном i-ом объекте вычисляется по формуле:

$x_i^{new} = \frac{x_i - x_{min}}{x_{max} - x_{min}}$, где x_{min}, x_{max} - минимальное и максимальное значения признака x соответственно. После такого преобразования минимальное значение каждого признака будет отображено в 0, а максимальное – в 1.

3. Робастная нормализация (Robust Scaling):

Данный способ преобразования похож на предыдущий, однако в нём для нормализации значений задействуется межквартильный диапазон, за счёт чего способ устойчив к выбросам (outliers), в отличие от первых двух рассмотренных преобразований, которые очень чувствительны к наличию выбросов в данных.

Для каждого признака x его новое значение на i-ом объекте может быть найдено следующим образом:

$x_i^{new} = \frac{x_i - Q_1(x)}{Q_3(x) - Q_1(x)}$, где $Q_1(x), Q_3(x)$ – первая (нижняя) и третья (верхняя) квартили признака x соответственно.

4. Квантильное преобразование (Quantile Transform)

Данное преобразование является нелинейным, новое значение признака x на i-ом объекте рассчитывается как:

$x_i^{new} = \Phi^{-1}(F(x_i))$, где Φ – функция распределения стандартного нормального закона, F – эмпирическая функция распределения. В результате признак будет принимать значения из диапазона [0,1]. Кроме того, также, как и робастная нормализация, квантильное преобразование устойчиво к выбросам (однако, для него даже выбросы будут переведены в значения из [0,1]).

5. Преобразование Йео-Джонсона (Yeo-Johnson power transformation)

Нелинейное преобразование Йео-Джонсона, описанное в [11], предназначено для придания данным более нормального распределения, стабилизации дисперсии, минимизации асимметрии признаков.

Для каждого признака x его значение на i-ом объекте преобразуется следующим образом:

$$x_i^{(\lambda)} = \begin{cases} \frac{((x_i+1)^{\lambda}-1)}{\lambda}, если \lambda \neq 0,\ x_i \geq 0 \\ \log(x_i+1), если\ \lambda = 0,\ x_i \geq 0 \\ \frac{-[(-x_i+1)^{(2-\lambda)}-1]}{2-\lambda}, если\ \lambda \neq 2,\ x_i < 0 \\ -log(-x_i+1), если \lambda = 2,\ x_i < 0 \end{cases}$$, где $0 \leq \lambda \leq 2$

При строго положительных значениях x_i оно представляет собой широко распространённое преобразование Бокса-Кокса (только для $x_i + 1$). Оптимальный параметр масштабирования λ для стабилизации дисперсии и минимизации асимметрии подбирается посредством максимизации функции правдоподобия независимо для каждого из признаков.

4.3. Восстановление пропущенных измерений фотоэффекта

В результате устранения шумов в исходных данных (2 методами) и их преобразования (5 способами) на предыдущих этапах получилось 10 новых наборов данных. Наконец, на каждом из этих 10 наборов будут проведены эксперименты с целью выявления наиболее эффективного способа решения задачи восстановления неизмеренных значений фотоэффекта для скважин ALEXANDER D, KIMZEY A и RECRUIT F9 методами машинного обучения.

Задача восстановления пропущенных значений признака на основе значений нескольких других признаков в терминах машинного обучения характеризуется как задача множественной регрессии. Она заключается в прогнозировании значений непрерывной целевой переменной, заданной на множестве вещественных чисел, на основе многомерных векторов признаков, характеризующих (т.е. задающих описание) исследуемые объекты. Пусть обучающая выборка состоит из N векторов признаков $\{x_n\}$ (где $n = 1, ..., N$) и соответствующих им целевых значений $\{t_n\}$. Необходимо для каждого нового вектора x из тестовой выборки спрогнозировать значение t целевой переменной с помощью построенной и обученной модели (условное распределение $P(t|x)$).

Конкретно для имеющейся задачи восстановления пропущенных измерений исследуемыми объектами являются литологические фации, в качестве признакового описания объектов были выбраны признаки GR, ILD_log10, DeltaPHI, PHIND, Depth из Таблицы 4 и также метка-класс фации (Facies), а целевой переменной, чьи значения необходимо прогнозировать, является фотоэффект (PE). Следует отметить, что класс фации является категориальным признаком, поэтому для корректной работы алгоритмов машинного обучения необходимо было предварительно применить к нему метод кодирования категориальных переменных в числовые (One-Hot Encoding).

Для решения поставленной регрессионной задачи были выбраны и опробованы следующие методы машинного обучения:

1. Метод k ближайших соседей (k-nearest neighbors - KNeighbors)

2. Адаптивный бустинг (adaptive boosting - Ada) - в качестве базовых алгоритмов использовались решающие деревья
3. Градиентный бустинг над решающими деревьями: использовалась его реализация XGBoost (экстремальный градиентный бустинг – extreme gradient boosting)
4. Случайный лес (random forest – RF)
5. Лассо-регрессия/ L1-регуляризация (least absolute shrinkage and selection operator – Lasso)
6. Ридж-регрессия или гребневая регрессия/ L2-регуляризация (ridge regression – Ridge)

Подробное описание каждого из вышеперечисленных методов представлено в работах [20], [34], [39], [21], [33] и [8] соответственно.

Каждый из методов был применен сначала к обучающему набору данных, а затем к тестовому. Стоит отметить, что в процессе обучения для каждого метода машинного обучения подбирались его основные параметры (процесс выбора модели – выбор оптимальной комбинации параметров) с целью повышения качества решения поставленной задачи. Подбор параметров осуществлялся по сетке – т.н. GridSearchCV [9], с помощью 7-блочной кросс-валидации.

Наиболее типичной мерой качества регрессионных моделей машинного обучения является среднеквадратичная ошибка (Mean Squared Error - MSE):

$MSE = \frac{1}{N}\sum_{i=1}^{N}(x_i - x_i^{pred})^2$, где N – количество наблюдений, x_i – истинное i-ое значение целевой переменной x, x_i^{pred} – предсказанное значение. Стоит отметить, что среднеквадратичный функционал сильно штрафует за большие отклонения от истинных значений, поэтому он чувствителен к выбросам в данных.

Тем не менее, MSE зависит от единиц измерения данных и скорее подходит для сравнения между собой нескольких моделей, но не позволяет сделать выводов о том, насколько эффективно конкретная модель решает поставленную задачу. В таком случае вместо среднеквадратичной ошибки полезно использовать коэффициент детерминации (R^2), который фактически является нормированной MSE и определяется следующим образом:

$R^2 = 1 - \frac{\sum_{i=1}^{N}(x_i - x_i^{pred})^2}{\sum_{i=1}^{N}(x_i - \bar{x})^2} = 1 - \frac{MSE}{\sigma_x^2}$, где $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$ - среднее значение целевой переменной. Коэффициент детерминации измеряет долю дисперсии,

объясненную моделью, в общей дисперсии целевой переменной. Чем ближе коэффициент к 1, тем лучше модель предсказывает значения целевой переменной.

Значения метрик R^2 и MSE в результате применения выбранных методов машинного обучения к данным, преобразованным различными способами (10 вариантов), для решения задачи восстановления значений фотоэффекта представлены в таблицах 6-9. Значения указаны для тестовой выборки, которая составляла 25 % от общего набора данных (рассматривались 7 скважин, для которых фотоэффект был измерен).

Метод / Преобразование	KNeighbors	Ada	**XGBoost**	RF	Lasso	Ridge
Стандартизация	0.822	0.619	0.835	0.73	0.357	0.669
Нормализация (Min-Max)	0.84	0.6	0.854	0.735	-0.0002	0.671
Робастная нормализация	0.843	0.625	0.856	0.744	0.364	0.654
Квантильное преобразование	0.837	0.634	**0.873**	0.761	-0.00002	0.687
Преобразование Йео-Джонсона	0.848	0.65	**0.868**	0.753	0.443	0.682

Таблица 6. Значения метрики R^2 различных методов машинного обучения для данных, обработанных методом EMD и преобразованных 5 способами.

Метод / Преобразование	KNeighbors	Ada	**XGBoost**	RF	Lasso	Ridge
Стандартизация	0.21	0.389	**0.16**	0.284	0.622	0.361
Нормализация (Min-Max)	0.002	0.004	**0.001**	0.003	0.01	0.004
Робастная нормализация	0.087	0.195	**0.08**	0.148	0.326	0.2
Квантильное преобразование	0.013	0.029	**0.01**	0.019	0.086	0.026
Преобразование Йео-Джонсона	0.146	0.336	**0.142**	0.248	0.54	0.317

Таблица 7. Значения метрики MSE различных методов машинного обучения для данных, обработанных методом EMD и преобразованных 5 способами.

Метод / Преобразование	KNeighbors	Ada	**XGBoost**	RF	Lasso	Ridge
Стандартизация	0.743	0.608	0.778	0.726	0.367	0.64
Нормализация (Min-Max)	0.718	0.604	0.774	0.72	-0.0001	0.616
Робастная нормализация	0.722	0.626	0.77	0.731	0.35	0.675
Квантильное преобразование	0.743	0.647	**0.79**	0.748	-0.0001	0.68
Преобразование Йео-Джонсона	0.764	0.626	**0.802**	0.752	0.436	0.664

Таблица 8. Значения метрики R^2 различных методов машинного обучения для данных, обработанных методом EEMD и преобразованных 5 способами.

Метод / Преобразование	KNeighbors	Ada	**XGBoost**	RF	Lasso	Ridge
Стандартизация	0.268	0.409	**0.228**	0.257	0.642	0.362
Нормализация (Min-Max)	0.004	0.005	**0.002**	0.003	0.012	0.004
Робастная нормализация	0.16	0.185	**0.135**	0.148	0.356	0.176
Квантильное преобразование	0.021	0.029	**0.017**	0.021	0.086	0.026
Преобразование Йео-Джонсона	0.223	0.39	**0.211**	0.244	0.54	0.316

Таблица 9. Значения метрики MSE различных методов машинного обучения для данных, обработанных методом EEMD и преобразованных 5 способами.

На основании полученных выше результатов можно сделать следующие выводы:

1. Эксперименты, проведенные на данных, обработанных с помощью метода EMD, оказались более успешными, чем эксперименты, осуществлённые на данных с EEMD-обработкой. Это ещё раз подтверждает выводы, полученные в разделе 4.1.
2. Самый высокий результат по совокупности значений рассматриваемых метрик показала модель экстремального градиентного бустинга (XGBoost), следующими по эффективности оказались методы KNeighbors и RF.
3. Можно отметить несостоятельность применения линейных моделей регрессии Lasso и Ridge к имеющимся данным со сложными связями между признаками, а также не очень высокую эффективность модели адаптивного бустинга (Ada).

4. В целом, регрессионные модели (в особенности, модели на основе XGBoost, RF и KNeighbors) показали себя более эффективными в применении к данным, обработанным с помощью нелинейных преобразований (квантильное и Йео-Джонсона).

На рисунках 9-10 приведены примеры восстановленных значений с помощью наиболее эффективной модели экстремального градиентного бустинга. Её оптимальные параметры, подобранные с помощью поиска по сетке, следующие: максимальная глубина дерева (max_depth) = 7, скорость обучения модели (learning_rate) = 0.1, а число решающих деревьев (n_estimators) = 300.

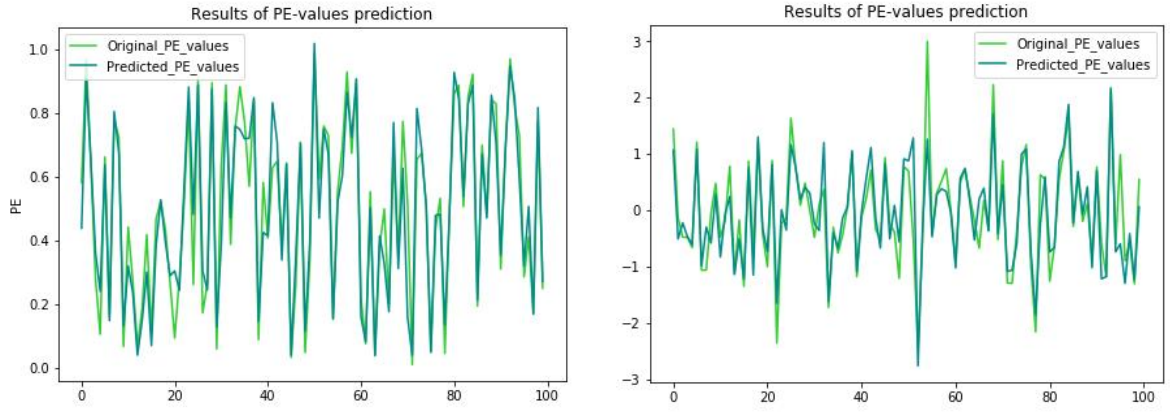

Рисунок 9. Результат восстановления значений фотоэффекта с применением XGBoost на данных, обработанных EMD, а также с помощью квантильного преобразования (слева) и преобразования Йео-Джонсона (справа).

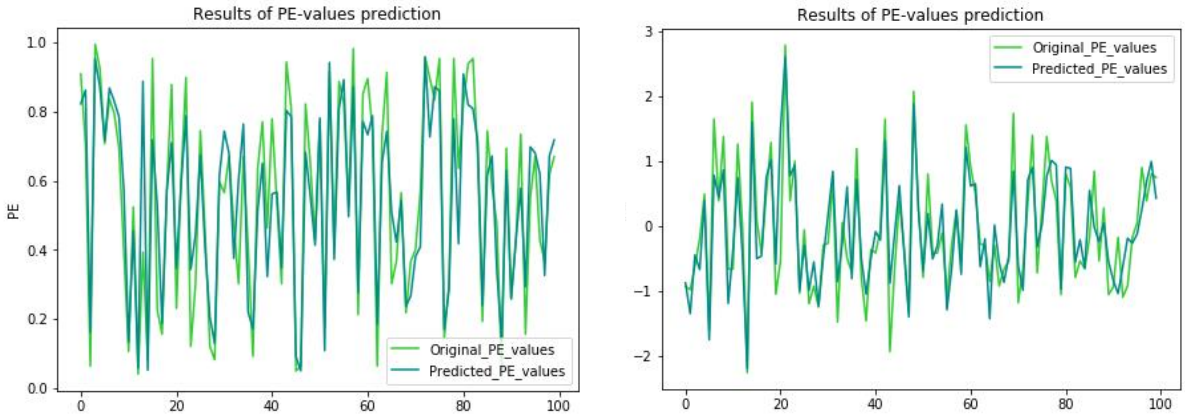

Рисунок 10. Результат восстановления значений фотоэффекта с применением XGBoost на данных, обработанных EEMD, а также с помощью квантильного преобразования (слева) и преобразования Йео-Джонсона (справа).

В итоге, для решения уже задачи классификации фаций будут задействованы данные, обработанные с помощью метода EMD, а также с помощью преобразования Йео-Джонсона, для которых значения фотоэффекта восстанавливались на основе применения метода XGBoost (R^2 составил 0.868).

Преобразование Йео-Джонсона оказалось предпочтительнее квантильного потому, что оно придает данным более нормальное распределение, предположение о котором зачастую необходимо для использования различных моделей машинного обучения (для проверки гипотезы нормальности распределения преобразованных данных использовался критерий Шапиро-Уилка [37]). Хотя и, в свою очередь, квантильное преобразование лучше работает с имеющимися выбросами в данных.

5. Задача классификации фаций

Данный раздел будет посвящен описанию завершающей ключевой стадии рабочего процесса, направленного на достижение главной цели исследования – поиска и осуществления эффективного способа классификации литологических фаций с помощью машинного обучения на основе пяти физических каротажных признаков GR, ILD_log10, DeltaPHI, PHIND и PE, которые были предварительно обработаны в разделе 4 настоящей работы, наряду с двумя геологическими признаками NM_M и RELPOS (Таблица 4).

Поставленная задача классификации фаций относится к одной из классических проблем в машинном обучении с учителем, а именно, к задаче мультиклассовой классификации (supervised multiclass classification problem), поскольку число фаций (т.е. классов из Таблицы 3) равно 9. Пусть обучающая выборка состоит из N примеров - векторов признаков $\{x_n\}$, $x_n \in R^k$, где $n = 1,...,N$ и $k = 7$ (т.к. имеется 7 признаков), а также из соответствующих им ответов – меток классов $\{y_n\}$, $y_n \in \{1,\ldots,9\}$. Требуется построить и обучить модель классификации H, решающую функцию, которая будет определять к какому из 9 классов относится каждый новый объект x уже из тестовой выборки: $H(x) = y$.

5.1. Метрики качества для многоклассовой классификации

Для оценки качества результатов мультиклассовой классификации прежде необходимо вспомнить, какие метрики используются для задач бинарной классификации (2 класса: 0 и 1), и обобщить их. Для перехода непосредственно к самим метрикам требуется ввести важное понятие для их описания в терминах ошибок классификации — матрицу ошибок (confusion matrix):

	y = 1	y = 0
$y_{pred} = 1$	TP	FP
$y_{pred} = 0$	FN	TN

Здесь y — истинная метка класса на объекте, y_{pred} — метка класса, предсказанная объекту моделью машинного обучения, TP — количество истинно-положительных предсказаний, FP — количество ложно-положительных предсказаний, FN — количество ложно-отрицательных предсказаний и TN — количество истинно-отрицательных предсказаний. Тогда отсюда возникает наиболее очевидная и распространённая метрика — доля правильных ответов (аккуратность), которая вычисляется как: $accuracy = \frac{TP + TN}{TP + TN + FP + FN}$. Однако, такая метрика является непоказательной

в случае, когда классы несбалансированные (Рисунок 11). Поэтому для оценки качества классификации принято также использовать метрики, которые не зависят от соотношения классов: $recall = \frac{TP}{TP + FN}$ — точность и $precision = \frac{TP}{TP + FP}$ — полнота. Метрика *recall* демонстрирует способность модели обнаруживать конкретный класс в принципе, а *precision* — способность отличать этот класс от остальных. Чтобы сохранить баланс между этими метриками, можно их объединить — рассчитать среднее гармоническое: $f_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$ (т.н. f_1-мера). Она достигает максимума в 1 при *recall* и *precision*, равных 1, и близка к 0, если хотя бы одна из метрик устремлена к 0.

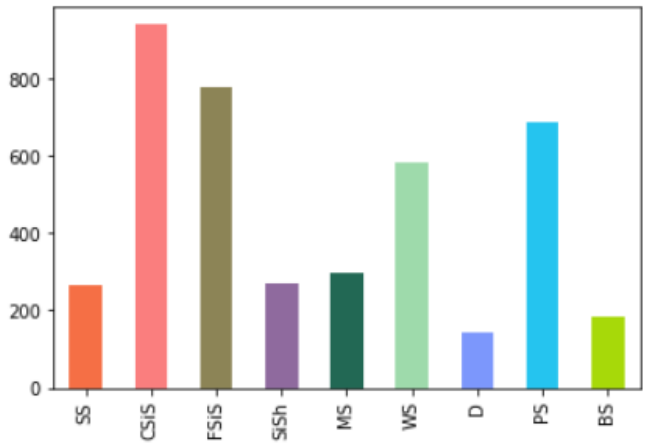

Рисунок 11. Число объектов для каждого из 9 имеющихся классов-фаций в исходном наборе данных (наблюдается дисбаланс классов).

Тогда, переходя к случаю, когда имеется 9 классов (т.е. рассматривается 9 задач бинарной классификации, каждая из которых отделяет один из классов от всех остальных), матрица ошибок будет выглядеть следующим образом:

	y = 1	…	y = 9
y_{pred} = 1	TP_1	…	FP_{19}
…	…	…	…
y_{pred} = 9	FN_{91}	…	TP_9

Для расчёта метрики *accuracy* по-прежнему необходимо оценить долю правильных ответов модели для всех 9 классов относительно общего числа ответов. В свою очередь, для расчёта *recall*, *precision* и, следовательно, для f_1-меры существует два подхода:

- Микро-усреднение. Для каждой из 9 задач вычисляются TP и FP, FN и TN, а затем на их основе вычисляются итоговые метрики следующим образом:

$$recall_{micro} = \frac{\sum_{i=1}^{9} TP_i}{\sum_{i=1}^{9} TP_i + \sum_{i=1}^{9} FP_i}, \quad precision_{micro} = \frac{\sum_{i=1}^{9} TP_i}{\sum_{i=1}^{9} TP_i + \sum_{i=1}^{9} FN_i}$$

  При таком способе усреднения вклад каждого класса в значения итоговых метрик зависит от его размера.
- Макро-усреднение. Вычисляются итоговые метрики для каждой из 9 задач, а затем усредняется по всем классам:

$$recall_{macro} = \frac{\sum_{i=1}^{9} recall_i}{9}, \quad precision_{macro} = \frac{\sum_{i=1}^{9} precision_i}{9}$$

  При таком способе усреднения все классы вносят вклад в равной степени.

Для оценки общих результатов поставленной задачи с помощью f_1-меры будет использоваться макро-усреднение. Кроме того, необходимо следить за значениями f_1-меры для каждой из 9 фаций в отдельности, чтобы выяснить, какие фации определяются моделью достаточно хорошо и какие, наоборот, выявляются плохо.

5.2. Методология

Для решения поставленной задачи мультиклассовой классификации литологических фаций были выбраны следующие методы машинного обучения:

- метод градиентного бустинга над решающими деревьями,
- случайный лес,
- глубокая свёрточная нейронная сеть
- метод k ближайших взвешенных соседей

Первые три подхода уже зарекомендовали себя как эффективные в применении к рассматриваемому набору данных каротажа – данный факт был установлен и описан в обзоре существующих решений (раздел 2, работы 4 периода [4, 46]). Последний же метод был выбран потому, что он показал достаточно высокие результаты при работе с имеющимся набором данных (который уже был предварительно очищен от шумов и преобразован) для решения задачи восстановления пропущенных измерений фотоэффекта, несмотря на свою простоту. На основании этого возникла идея, что данный метод может также показать неплохие результаты для задачи классификации фаций.

Обзор каждого из вышеобозначенных методов будет представлен ниже, кроме того, для каждого из них будут настроены основные параметры (т.е. осуществлён выбор модели), однако, для свёрточной нейронной сети в данной работе будет использована её уже готовая архитектура (модель с уже настроенными гиперпараметрами), предложенная авторами исследования [46]. Соответственно, основной целью применения этой свёрточной нейронной сети является попытка воспроизвести эксперименты и результаты, полученные в [46].

5.2.1. Градиентный бустинг над решающими деревьями

Градиентный бустинг над решающими деревьями – это метод машинного обучения, позволяющий для увеличения точности предсказания строить модель классификации в виде композиции базовых классификационных моделей, а именно, деревьев решений [7]. В качестве итоговой вероятности принадлежности какого-либо объекта конкретному классу выступает взвешенная сумма вероятностей принадлежности этого объекта классу каждого из классификаторов в композиции.

Выстраиваемая модель обучается последовательно – на каждой итерации происходит наращивание композиции классификаторов с целью уменьшения среднеквадратичного отклонения (MSE) предсказаний модели (т.е. каждый раз, при добавлении новой базовой модели решающего дерева в композицию, его структура определяется с учётом среднеквадратичного отклонения предсказательной модели с предыдущего этапа): $c_N(x) = c_{N-1}(x) + \alpha_N b(x, \beta_N)$, где $c_{N-1}(x)$ – композиция, построенная на *N-1* шаге, $b(x, \beta_N)$ – базовый классификатор, т.е. решающее дерево с параметрами β_N, α_N – коэффициент взвешенной суммы. Коэффициенты, в свою очередь, постепенно переподбираются так, чтобы оптимизировать (т.е. минимизировать) уже не MSE, а некоторую выбираемую заранее функцию потерь с помощью метода градиентного спуска. В результате такого подхода скорость сходимости градиентного бустинга увеличивается в значительной степени.

Существует несколько различных реализаций алгоритма градиентного бустинга над решающими деревьями, в данной работе будет использоваться его оптимизированный вариант – XGBoost (экстремальный градиентный бустинг). На сегодняшний день, XGBoost является самой эффективной реализацией из всех. Оптимизация достигается за счёт того, что рассматриваемая функция для оптимизации (сумма функций потерь с добавлением регуляризации для борьбы с переобучением) градиентного бустинга приближается некоторым выражением с помощью разложения

Тейлора до второго члена. Подробное описание всех особенностей XGBoost представлено в работе [39].

Для улучшения метрик качества многоклассовой классификации, осуществляемой с помощью XGBoost, его следующие основные параметры были выбраны для настройки: число решающих деревьев (n_estimators), максимальная глубина дерева (max_depth) и скорость обучения (learning_rate). Что касается параметра max_depth, при его определении важно сохранять баланс: слишком неглубокие деревья не способны фиксировать важные особенности, присущие набору данных, в свою очередь, слишком глубокие деревья склонны к потере обобщающей способности, поскольку они фиксируют специфичные особенности, характерные для конкретной выборки. Параметр learning_rate отвечает за устойчивость модели, предотвращает её переобучение (посредством сокращения размера шага).

Лучшая комбинация параметров XGBoost была подобрана с помощью поиска по сетке (при участии 5-блочной кросс-валидации) для n_estimators в диапазоне от 100 до 450 с шагом 50, max_depth в диапазоне от 3 до 18 с шагом 1, learning_rate в диапазоне от 0.1 до 1.0 с шагом 0.1.В качестве функции потерь используется логарифмическая функция потерь для мультиклассовой классификации. На рисунке 12 представлена зависимость метрики *accuracy* (её значения указаны для классификации объектов из тестовой выборки, которая составляла 20% от общего набора данных) от параметров max_depth и n_estimators рассматриваемого метода при скорости обучения, равной 0.2.

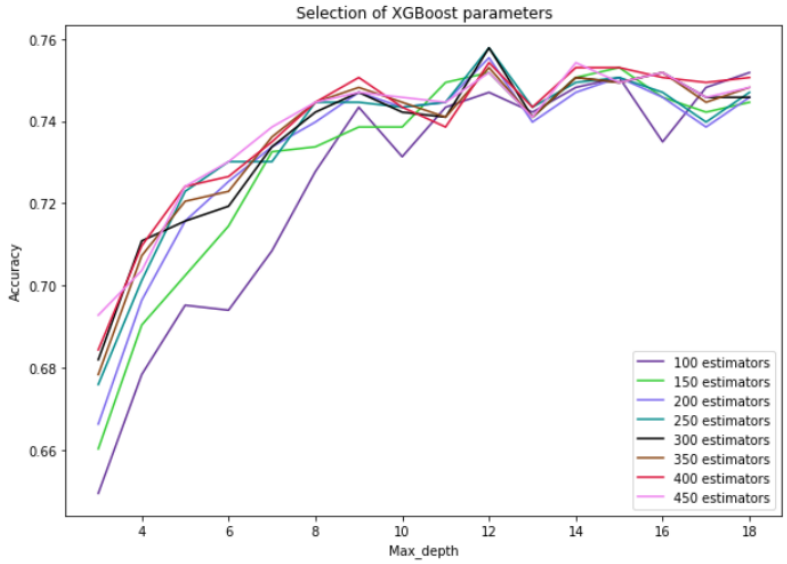

Рисунок 12. Зависимость значений метрики *accuracy* от комбинации параметров *XGBoost*.

Самую высокое значение метрики *accuracy* (75.8 %) для классификации объектов из тестовой выборки метод показал при 300 деревьях, максимальной

глубине 12 и скорости обучения, равной 0.2 (f_1-мера составила 74.1 %). Классификационная модель с данными параметрами далее будет использована для проведения ряда экспериментов (раздел 5.3).

5.2.2. Случайный лес

Данный алгоритм машинного обучения, основной принцип которого заключается в использовании ансамбля решающих деревьев для снижения такого явления, как переобучение, и повышения точности классификации (в сравнении с использованием лишь одного решающего дерева), описан в работе [21]. Случайный лес (Random Forest) сочетает в себе сразу две ключевые идеи: бэггинг (Bootstrap AGgregrating – bagging), предложенный Лео Брэйманом, и метод случайных подпространств (random subspace method - RSM). Бэггинг подразумевает независимое обучение каждого классификатора-решающего дерева, а также независимое предсказание класса конкретному объекту из тестовой выборки каждым из классификаторов и определение итогового результата для этого объекта путём голосования (итоговым классом будет тот, за который проголосовало большинство классификаторов, при условии, что одно дерево обладает лишь одним голосом). При этом, на основе метода случайных подпространств, каждый классификатор обучается на некотором подмножестве (разбиении) обучающей выборки, выделяемом случайный образом с повторением. Кроме того, при обучении в ходе создания очередного узла решающего дерева, выбор признака, на основе которого осуществляется разбиение, происходит не среди всех имеющихся признаков, а лишь среди r случайно выбранных, что позволяет строить более разнообразные деревья. Классически, выбор одного наилучшего признака для разбиения из r осуществляется с помощью критерия Джини (Gini criterion) [22].

Для наиболее эффективной работы алгоритма Random Forest в применении к имеющемуся набору данных, его следующие основные параметры было необходимо оптимизировать: число решающих деревьев в лесу (n_estimators), максимальная глубина дерева (max_depth) и критерий выбора признака для разбиения (criterion). Поиск оптимальной комбинации параметров был осуществлен с помощью стратегии поиска по сетке для n_estimators ∈ {50, 80, 100, 150, 200, 250, 300}, max_depth в диапазоне от 5 до 18 с шагом 1 и criterion ∈ {критерий Джини, критерий прироста информации (entropy)}. Зависимость доли правильных ответов при классификации объектов из тестовой выборки (20% от общего набора данных) от параметров n_estimators и max_depth рассматриваемого метода с критерием выбора признака, установленном как критерий Джини, представлена на рисунке 13.

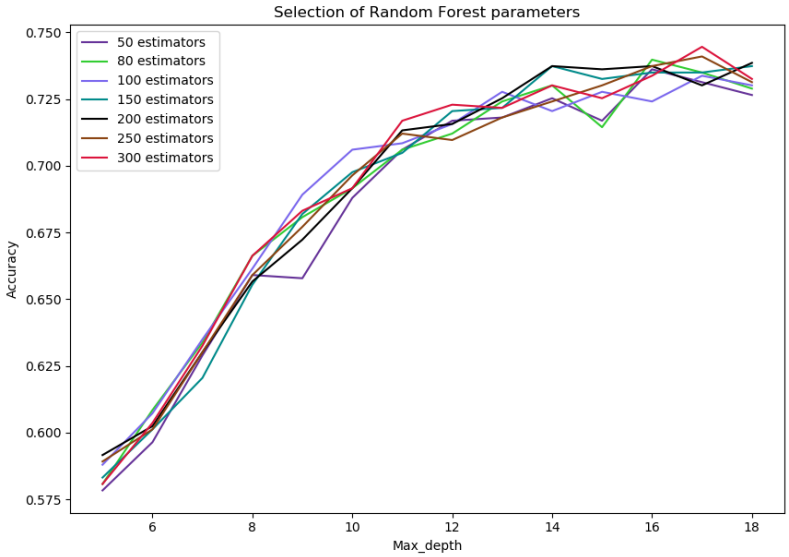

Рисунок 13. Зависимость значений метрики *accuracy* от комбинации параметров метода *Random Forest*.

На основании полученных результатов можно сделать вывод, что наивысшее значение метрики *accuracy* в 73.8 % для классификации объектов из тестовой выборки *Random Forest* достигает при 300 деревьях, максимальной глубине 17 и критерии Джини выбора признака (f_1-мера составила 73 %). В разделе 5.3 настоящего исследования полученная классификационная модель будет использована для проведения ряда экспериментов.

5.2.3. Метод *k* взвешенных ближайших соседей

Данный метод, описанный в [29, 42], является одним из наиболее известных метрических алгоритмов для автоматической классификации объектов. Его ключевая идея основана на измерении степени сходства между объектами с помощью определённо заданной метрики расстояния и предположении о том, что близкие друг к другу объекты находятся в одном классе. Каждый объект x из тестовой выборки классифицируется путём голосования по его *k* ближайшим соседям из обучающей выборки. Каждый из *k* соседей голосует за отнесение объекта x к своему классу, в итоге метод присваивает объекту тот класс, который набирает наибольшее число голосов. Однако, может возникнуть ситуация, в которой максимальная сумма голосов достигается сразу на нескольких классах, поэтому необходимо для каждого соседа объекта задать вес – его степень влияния на результат классификации (самый большой вес назначается соседу, являющемуся ближайшим к классифицируемому объекту, самый маленький – самому дальнему из *k* соседей).

Наиболее значимыми параметрами, которые необходимо оптимизировать для наиболее эффективной работы алгоритма, являются число соседей и мера расстояния между объектами. Число соседей для поставленной задачи рассматривалось в диапазоне от 2 до 20. В качестве мер расстояния были рассмотрены следующие:

- Евклидово расстояние: $\sqrt{\sum_{i=1}^{7}(x_i - y_i)^2}$ для объектов x и $y \in R^7$ (поскольку классифицируемые объекты в рассматриваемой задаче описываются 7 признаками)
- Расстояние городских кварталов (манхэттенское): $\sum_{i=1}^{7} | x_i - y_i |$
- Расстояние Чебышёва: $\max_{i=1,\dots,7} | x_i - y_i |$

Выбор оптимальной комбинации параметров вида (число соседей, мера) осуществлялся с помощью стратегии поиска по сетке (при участии 5-блочной кросс-валидации). В качестве весов для соседей использовались инвертированные расстояния от них до классифицируемого объекта. Доля правильных ответов (*accuracy*) при классификации объектов из тестовой выборки, которая составляла 20% от общего набора данных, в зависимости от параметров метода представлена на рисунке 14.

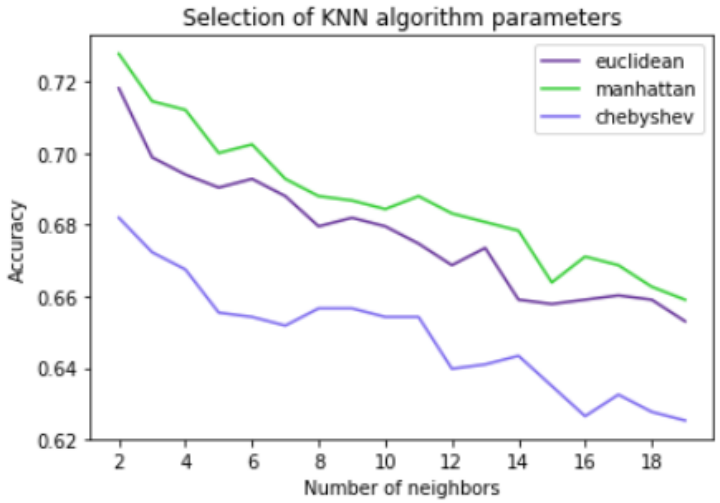

Рисунок 14. Зависимость значений метрики *accuracy* от комбинации параметров метода *k ближайших соседей*.

Так, самое высокое значение *accuracy* в 73.1 % для классификации объектов из тестовой выборки метод показал при числе соседей, равному 2, с вычислением расстояний между объектами с помощью манхэттенской метрики (f_1-мера составила 72.5 %). Данная модель классификации в

дальнейшем будет использована для проведения ряда экспериментов (раздел 5.3).

5.2.4. Одномерная свёрточная нейронная сеть

На сегодняшний день, развитие разнообразных техник машинного обучения привело к широкому распространению искусственных нейронных сетей (ИНС) — систем соединённых и взаимодействующих между собой искусственных нейронов (упрощенные модели биологических нейронов). ИНС фактически являются машинной интерпретацией работы человеческого мозга, который состоит из миллиардов клеток – нейронов, накапливающих и передающих информацию в виде электрических импульсов. Так же и ИНС обладают такой структурой, которая позволяет анализировать, запоминать и даже воспроизводить в памяти различную информацию. Эти особенности позволяют использовать ИНС для решения самых различных задач, например, классификации, прогнозирования, распознавания образов.

Свёрточная нейронная сеть (*convolutional neural network - CNN*) — специальная архитектура нейронных сетей, относящихся к классу глубокого машинного обучения [16], которая была предложена Я. Лекуном [47], изначально нацеленная на эффективное решение задач компьютерного зрения. *CNN* состоит из трёх основных видов слоёв: свёрточный слой (convolutional layer), слой субдискретизации – подвыборки (pooling layer) и выходной полносвязный слой (fully-connected layer). Принцип работы *CNN* нагляднее всего можно описать в случае, когда стоит задача обработки двумерных изображений (принцип работы одномерной *CNN* аналогичен, только на её вход вместо изображения подаётся вектор признаков). Так, на свёрточном слое к разным подматрицам изображения (двумерная матрица) применяется математическая операция свёртки (фильтр) [41]. Результатом этих действий является множество вещественных чисел – т.н. сформированная карта признаков (осуществляется переход от изображения в целом к наиболее существенным его деталям, маловажные детали отфильтровываются). На одном свёрточном слое может применяться несколько фильтров (их количество - гиперпараметр), тогда число сформированных карт признаков будет равно их числу.

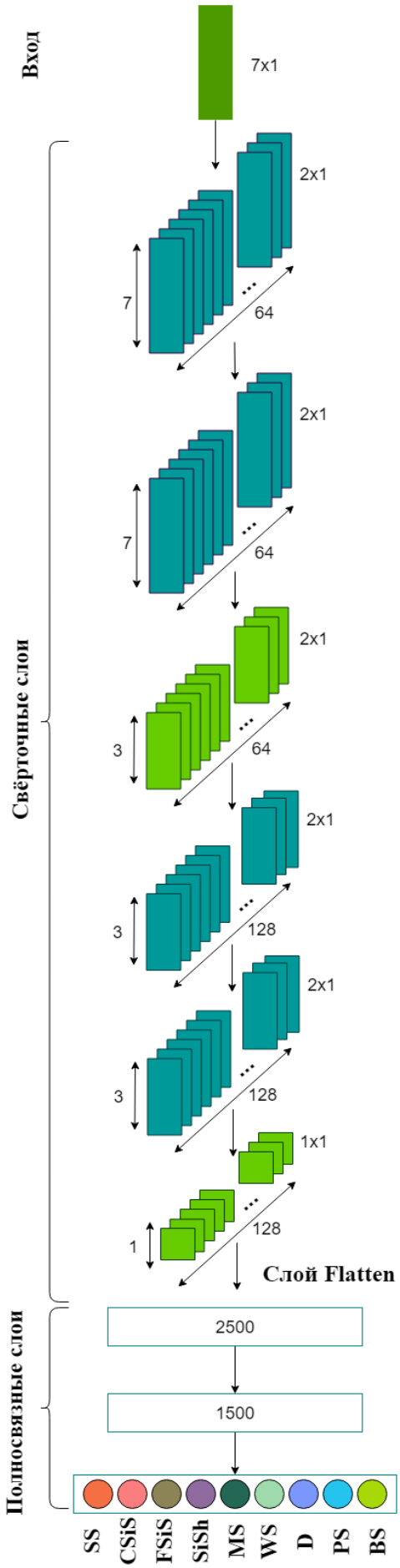

Рисунок 15. Архитектура одномерной свёрточной нейронной сети, предложенной в исследовании [46]. Для каждого свёрточного и субдискретизирующего слоя указаны следующие гиперпараметры: слева – размерность признакового пространства (после применения слоя субдискретизации размерность снижается с 7 до 3), посередине – количество фильтров, справа – длина фильтра (размер ядра свёртки). Для последующих трёх полносвязных слоёв указано число нейронов на них – 2500, 1500 и 9 (по числу классов) соответственно.

Далее субдискретизация применяется к результатам свёртки для их преобразования в меньшее количество элементов (уменьшение карты признаков) и увеличения степени инвариантности применяемых фильтров. Так, например, max-pooling субдискретизация предоставляет локальные максимальные значения из результатов свёртки на каких-то их подмножествах (для соседей), а mean-pooling находит среднее арифметическое. Такая операция позволяет уменьшить вычислительную сложность и снизить вероятность переобучения нейронной сети. Таким образом, *CNN* строится путём чередования двух видов слоёв – свёртки и субдискретизации. На выходе располагаются полносвязные слои; каждый нейрон полносвязного слоя представляет собой перцептрон с нелинейной функцией активации, который соединяется с выходами нейронов предыдущего слоя. С помощью *softmax*-функции (обобщение логистической функции на многомерный случай) выдаются вероятности принадлежности входного объекта к каждому из имеющихся классов.

Авторы исследования [46] для решения задачи классификации фаций использовали нейронные сети сразу нескольких типов – рекуррентную нейронную сеть (*RNN*), нейронную сеть с долгой кратковременной памятью (*LSTM*) и одномерную свёрточную нейронную сеть (*1D-CNN*), и пришли к выводу, что *1D-CNN* показывает наилучшие результаты. В качестве возможной мотивации применения авторами нейронной сети, которая обычно используется для распознавания изображений (двумерный случай), может выступать тот факт, что одномерные *CNN* сейчас активно используются для решения задач обработки последовательностей и поиска в них шаблонных подпоследовательностей, в частности, для задач обработки естественного языка [43] (есть некая аналогия с поставленной задачей классификации: важно учитывать контекст, т.е. глубину залегания фаций, поскольку можно наблюдать некоторый шаблон, цикличность их залегания – стратиграфическая информация о циклическом чередовании фаций [5]).

Архитектура *1D-CNN*, предложенная в [46], приведена на рисунке 15. Она состоит из входного слоя, на который подавался 7-мерный вектор значений каротажных измерений, четырёх свёрточных слоев с усечённым линейным преобразованием (rectified linear unit - ReLU) в качестве функции активации, двух слоев максимальной субдискретизации (max-pooling) для уменьшения вычислительной нагрузки и трёх полносвязных слоёв. На последнем выходном слое *1D-CNN* входным данным (объектам) назначается фациальная метка (класс). Обучение нейронной сети было направлено на поиск наилучших параметров (т.е. весов нейронных связей) для минимизации функции потерь, которая в поставленной задаче измеряла совместимость между прогнозируемым значением класса и истинной фациальной меткой. В

качестве функции потерь была выбрана категориальная кросс-энтропия, определяющая, насколько прогнозируемое распределение вероятностей принадлежности исследуемого объекта к выделенным классам близко к истинному распределению. Все веса и смещения нейронов модели оптимально настраивались с помощью метода адаптивного градиента *Adagrad*, минимизирующего функцию потерь, то есть использующего кросс-энтропию в качестве обучающей метрики. Кроме того, дополнительный настроенный механизм прореживания (dropout) был применен авторами ко второму и четвертому свёрточным слоям, а также к первому полносвязному слою *1D-CNN*, чтобы избежать проблемы переобучения модели. Так, dropout отбросил некоторые случайно выбранные скрытые нейроны, и они не использовались в *1D-CNN* на стадии обратного распространения. Также авторами был применен метод Batch Normalization после каждого свёрточного и первого полносвязного слоёв нейросети для того, чтобы увеличить скорость её обучения и сделать ее менее чувствительной к инициализации. Гиперпараметр batch size (количество примеров в одном обучающем проходе *1D-CNN*) был определён как 10, а число эпох (прогонка всех обучающих примеров через нейронную сеть в обоих направлениях) — 4000.

Вышеописанная модель *1D-CNN* была имплементирована с гиперпараметрами (параметры, значения которых задаются до начала обучения модели и не изменяются в его процессе), которые были подобраны и указаны авторами в их работе [46]. Однако, в ходе исследования было установлено, что batch size, равный 64, и число эпох, равное 2800, являются более оптимальными значениями данных гиперпараметров (нежели 10 и 4000 соответственно) с точки зрения эффективности и точности модели *1D-CNN*. Далее было осуществлено обучение модели с целью оптимизации её параметров (т.е. уже весов нейронных связей и смещений) так, чтобы минимизировать функцию потерь. На рисунке 16 представлен процесс обучения полученной модели (данные для обучения составили 80 % от общего набора данных). Обученная модель далее будет использована для классификации объектов из выборки, не участвовавшей в обучении (эксперименты в разделе 5.3).

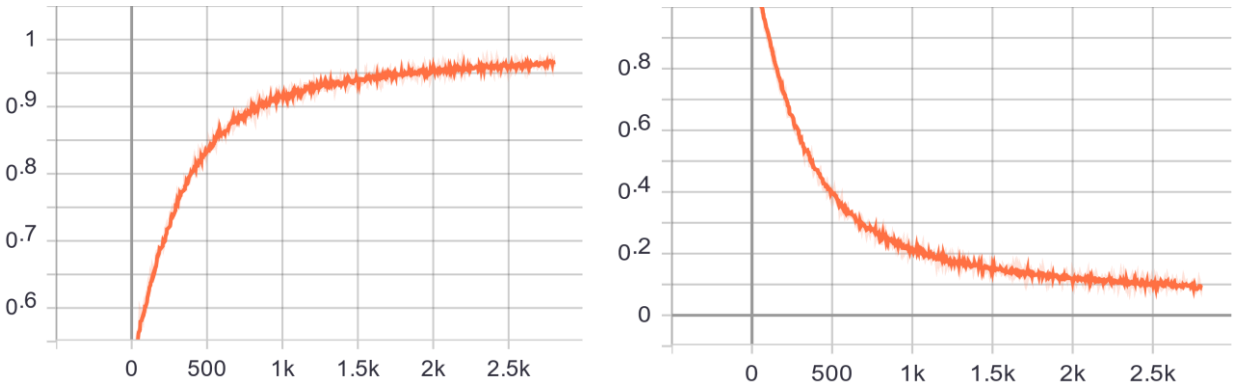

Рисунок 16. На графике слева представлено изменение значений метрики *accuracy* в процессе обучения (эпохи), на графике справа отражено поведение функции потерь (для обучающей выборки)

5.3. Эксперименты

Настроенные на предыдущем этапе модели машинного обучения для решения задачи классификации фаций были применены как к предварительно обработанному в разделе 4 набору данных каротажа, где устранялись шумы, а пропущенные измерения восстанавливались с помощью градиентного бустинга над решающими деревьями (*data set 1*), так и к исходному набору данных с минимальной предобработкой (*data set 2*). Под минимальной предобработкой подразумевается использование подхода, применяемого авторами работ, которые были описаны в разделе 2 (работы 4 периода). Данный подход заключался в заполнении пропущенных измерений фотоэффекта средним значением по всем скважинам, для которых он был известен, а также в стандартизации признаков. Это необходимо, поскольку позволит оценить целесообразность и эффективность произведённой работы по обработке каротажных измерений (т.е. позволит проверить, каким образом предобработка повлияет на качество классификации).

Кроме того, при проведении экспериментов на обоих наборах данных, была использована кросс-валидация (метод перекрёстной проверки), поскольку важно было проверить и оценить достоверное качество, обобщающую способность каждой построенной модели, а также осуществить объективное сравнение их эффективности. Для этих целей была выбрана k -блочная (*k-fold*) стратегия кросс-валидации (число блоков $k = 5$), которая заключается в следующем: исходный набор данных разбивается на k непересекающихся, примерно одинаковых по объёму частей, далее осуществляется k итераций, на каждой из которых один из полученных блоков становится тестовой выборкой (на которой осуществляется оценка качества модели), а остальные $k-1$ блоков используются для обучения модели. Каждый из k блоков становится выборкой

для тестирования единожды. Соответственно, при 5-блочной стратегии, на каждой итерации исходный набор данных разбивается в соотношении 80% (3319 объектов) к 20 % (830 объектов) на обучающую и тестовую выборки. В результате, получается *k* оценок, по каждой на один блок; чтобы получить итоговую оценку, необходимо усреднить эти оценки. Применение такого подхода позволило наиболее равномерно использовать имеющиеся данные.

В таблицах 10 и 11 приведены значения метрик качества имеющихся классификационных моделей машинного обучения, полученные в результате применения моделей к *data set 1* и *data set 2* соответственно (значения указаны на основе результатов кросс-валидации). Кроме того, на рисунке 17 для каждой из 4 моделей представлено сравнение между собой её значений метрик качества, полученных для двух наборов данных, предобработанных различными способами.

Метод / Метрика	Градиентный бустинг над решающими деревьями (XGBoost)	Случайный лес (Random Forest)	Метод *k* ближайших взвешенных соседей (k-NN)	Одномерная свёрточная нейронная сеть (1D-CNN)
accuracy	75.01 %	73.48 %	72.88 %	**77.22 %**
f_1-мера	73.85 %	72.89 %	72.47 %	**75.36 %**

Таблица 10. Значения метрик качества классификации объектов из тестовой выборки (*data set 1)* для различных моделей машинного обучения

Метод / Метрика	Градиентный бустинг над решающими деревьями (XGBoost)	Случайный лес (Random Forest)	Метод *k* ближайших взвешенных соседей (k-NN)	Одномерная свёрточная нейронная сеть (1D-CNN)
accuracy	72.08 %	70.6 %	67.87 %	**74.81 %**
f_1-мера	71.43 %	69.7 %	66.54 %	**74.2 %**

Таблица 11. Значения метрик качества классификации объектов из тестовой выборки (*data set 2)* для различных моделей машинного обучения

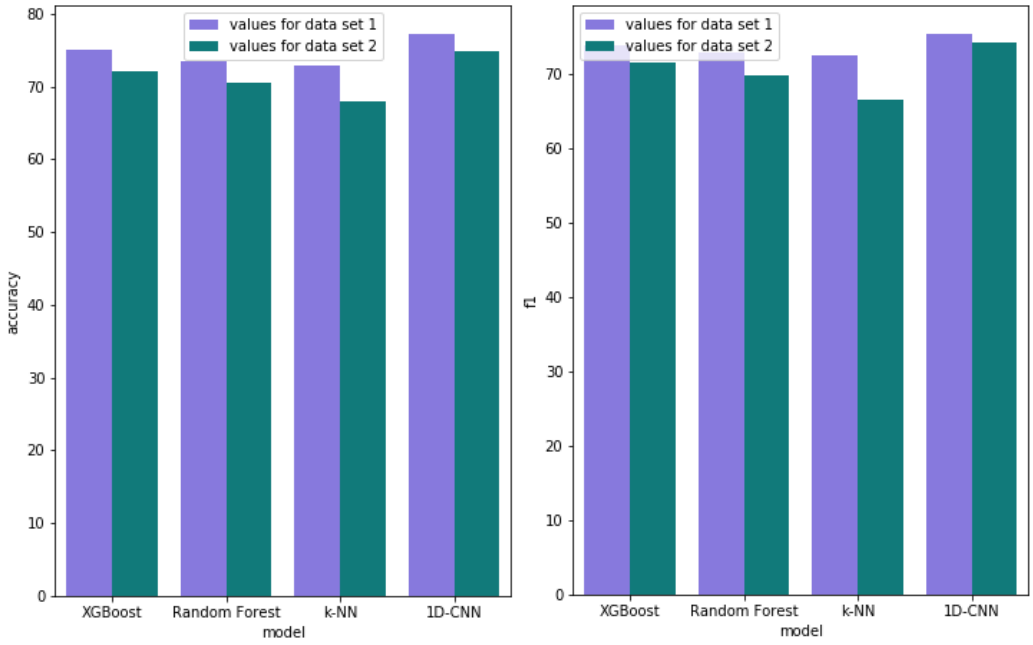

Рисунок 17. Слева для каждой имеющейся модели сравниваются значения *accuracy*, полученные при проведении экспериментов на *data set 1* и *data set 2*, справа – значения f_1-*меры*

На основании указанных выше результатов можно сделать следующие выводы:

- Для всех построенных классификационных моделей значения метрик оказались в среднем на 3-4 % выше в результате проведения экспериментов на первом предварительно обработанном наборе данных каротажа (*data set 1*), чем на наборе данных с минимальной предобработкой (*data set 2*). Таким образом, можно заключить, что выдвинутое ранее предположение о целесообразности осуществления предварительной обработки данных геофизических исследований скважин для решения задачи определения фаций подтверждается полученными результатами экспериментов. Кроме того, стоит отметить, что на результаты классификации для модели на основе более наивного подхода - метода *k-NN* (в сравнении с другими рассматриваемыми методами машинного обучения) наличие тщательной предобработки повлияло в самой значительной степени: значения метрик повысились примерно на 5-6 %. Это можно объяснить тем, что метод *k-NN* менее устойчив к наличию различных особенностей (недостатков) в данных, чем остальные методы (например, *XGBoost* и *Random Forest* изначально не требуют преобразования данных и даже имеют внутренние механизмы восстановления пропусков).

- Каждая из рассматриваемых моделей машинного обучения в целом оказалась приемлимо эффективной с точки зрения значений выбранных

метрик качества: так, для первого набора данных (с предобработкой, предложенной в разделе 4 данной работы) *accuracy* принимает значения свыше 72%, f_1-мера – также свыше 72 %, а для второго набора данных с минимальной предобработкой *accuracy* составляет не ниже 67%, f_1-мера – не ниже 66 %.

- Наиболее эффективной по совокупности рассматриваемых метрик качества оказалась модель одномерной свёрточной нейронной сети *1D-CNN* (для обоих наборов данных): для *data set 1 accuracy составила 77.22 %, а* f_1-мера – 75.36 %, для *data set 2* – 74.81 % и 74.2 % соответственно. Отсюда следует сразу два итога: во-первых, в данном исследовании в целом удалось воспроизвести эксперименты и результаты работы [46], во-вторых, удалось подтвердить целесообразность применения техник глубокого обучения к поставленной задаче интерпретации данных геофизической разведки.

- Однако, при сравнении эффективности моделей для решения поставленной задачи также стоит принять во внимание скорость их обучения, степень сложности подбора оптимальных параметров: так, градиентный бустинг над решающими деревьями и свёрточная нейронная сеть значительно сложнее в настройке (в особенности *CNN*), а модели на их основе требуют большего времени на обучение, в сравнении с случайным лесом и, тем более, с методом *k* ближайших соседей. В условиях ограниченности ресурсов использование моделей *Random Forest* и *k-NN* может оказаться предпочтительнее.

- Сравнивая полученные результаты с результатами других работ из обзора [3,4,46], посвящённых поставленной задаче (раздел 2, таблица 1), разработанный в данной работе способ решения задачи классификации литологических фаций на основе данных каротажных измерений (включающий в себя предварительную обработку данных, выбор методов машинного обучения и настройку их параметров) можно считать весьма конкурентоспособным.

Как упоминалось ранее, важно также следить за значениями f_1-меры для каждой из 9 фаций (классов) в отдельности, поскольку зачастую при разработке и эксплуатации скважин месторождения выделяются наиболее важные (ключевые) фации, высокоэффективное определение которых является необходимым и более приоритетным, чем знание о том, насколько хорошо определяются все фации в целом. На рисунке 18 для всех 9 имеющихся классов указаны значения f_1-меры, полученные в результате применения модели *1D-CNN* для классификации объектов из тестовой выборки набора

данных с предварительной предобработкой *(data set 1)*. Эти значения находятся в диапазоне от 65.3 % до 82.87 %. Несколько хуже остальных классов определились фации FSiS и SiSh (классы из таблицы 3) – 67.21 % и 65.3 % соответственно (хотя значение метрики для конкретного класса зависит и от того, в каком соотношении объекты этого класса попадают в обучающую и тестовые выборки, а также от общего количества объектов класса в наборе данных). Для всех других рассмотренных моделей машинного обучения значения f_1-меры для каждой фации не опускались ниже 60 %.

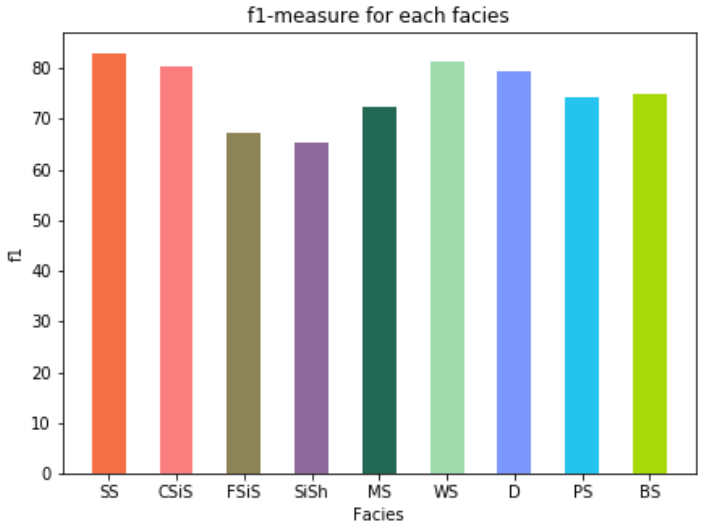

Рисунок 18. Значения f_1-меры для каждой из 9 фаций

Наконец, необходимо рассчитать долю правильных ответов (*accuracy*) для задачи классификации фаций с учётом смежных. В разделе описания исходного набора данных указывалась его следующая особенность: некоторые классы-фации являются смежными (Таблица 3) в том смысле, что они схожи по своим петрофизическим свойствам (такие фации сложнее различать). Поэтому наличие классификации, близкой к фактической (т.е. с учётом смежных фаций), в какой-то мере можно считать удовлетворительным решением. Так, например, если какой-то объект из тестовой выборки в результате классификации относится не к своему истинному классу, но к классу, который является смежным с ним, то такое решение классификатора не будет считаться ошибочным. На рисунке 19 приведены значения метрики *accuracy* для классификации фаций из тестовой выборки предобработанного набора данных с учётом смежных с помощью имеющихся моделей машинного обучения: *XGBoost* – 94.69 %, *Random Forest* – 93.37 %, *k-NN* – 92.89 % и *1D-CNN* – 95.25 %. На основании этих результатов можно сделать вывод о практически полном отсутствии слишком грубых ошибок (неверных классификаций) при определении фаций каждой из моделей.

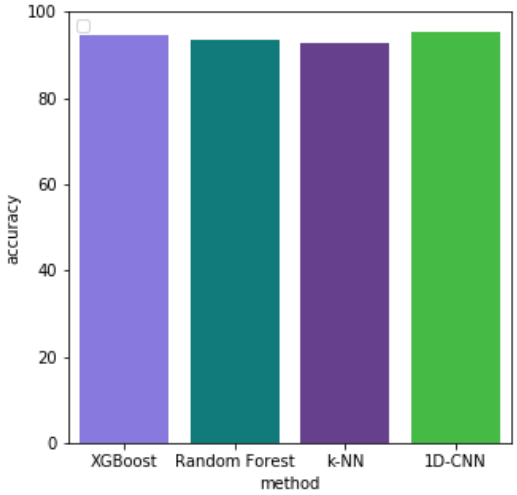

Рисунок 19. Значения метрики *accuracy* для задачи классификации фаций с учётом смежных

Заключение

В рамках данного исследования была поставлена цель разработки эффективного способа классификации литологических фаций на основе данных каротажа, полученных в ходе проведения геофизических исследований скважин, методами машинного обучения. Для достижения заявленной цели, в ходе работы был выполнен ряд важных задач.

Во-первых, был осуществлён подробный обзор существующих решений проблемы автоматической классификации фаций за последние 30 лет. Во-вторых, был найден удовлетворяющий цели исследования набор данных геофизических исследований скважин месторождения Панома, который был детально изучен и описан с указанием всех присущих ему особенностей.

В-третьих, поскольку для решения задачи классификации фаций было принято решение использовать методы машинного и, в частности, глубокого обучений, в работе особое внимание уделялось изучению способов предварительной обработки исходных данных каротажа с целью повышения качества и эффективности работы выбранных методов. Так, была осуществлена 3-х этапная предобработка исходного набора данных с учётом его выявленных ключевых особенностей, которая заключалась в следующем: сначала каротажные измерения отчищались от шумов с помощью методов эмпирической модовой декомпозиции и ансамблевой эмпирической декомпозиции (сравнительное исследование показало, что метод эмпирической модовой декомпозиции эффективнее справляется с задачей), затем данные были преобразованы, и, наконец, пропущенные значения измерений фотоэффекта были восстановлены с помощью следующих методов машинного обучения: метод k ближайших соседей, адаптивный бустинг, градиентный бустинг над решающими деревьями (реализация XGBoost), случайный лес, Лассо-регрессия и Ридж-регрессия. В итоге, для восстановления пропусков использовался XGBoost (с предварительно подобранными оптимальными параметрами), поскольку он показал наилучший результат (R^2 составил 0.873). Стоит отметить, что ранее ни в одной из работ по схожей тематике не уделялось внимание тщательной предобработке исходных данных, поэтому можно считать осуществление таковой главной отличительной чертой данного исследования.

Далее, на основе произведённого обзора, для решения уже задачи классификации фаций были выбраны следующие методы машинного обучения: градиентный бустинг над решающими деревьями, случайный лес, глубокая одномерная свёрточная нейронная сеть и метод k ближайших соседей. Кроме того, были выбраны метрики качества многоклассовой

классификации (accuracy, f_1-мера с макро-усреднением), а также была произведена настройка основных параметров выбранных методов (выбор моделей). Затем, построенные классификационные модели были использованы для проведения экспериментов сразу на двух наборах данных – как на предварительно обработанном, так и на исходном. Так, было установлено, что для всех моделей значения выбранных метрик оказались в среднем на 3-4 % выше в результате проведения экспериментов на данных с предобработкой. Таким образом, выдвинутое предположение о целесообразности осуществления предварительной обработки данных геофизических исследований скважин для решения задачи определения фаций было подтверждено.

Кроме того, при сравнении классификационных моделей между собой, было выяснено, что наиболее эффективной по совокупности рассматриваемых метрик является модель одномерной свёрточной нейронной сети: accuracy составила 77.22 %, f_1-мера – 75.36 %. Это подтвердило целесообразность применения техник глубокого обучения к задаче интерпретации данных геофизической разведки (классификации фаций). Наконец, сравнивая полученные результаты с результатами других исследований, можно сделать вывод, что предложенный способ решения задачи классификации фаций, включающий в себя предобработку данных и настройку параметров методов, является одним из самых конкурентоспособных.

На основании всего вышеперечисленного поставленную в данной работе цель можно считать достигнутой. Исходный код реализации разработанного способа находится по адресу: https://github.com/JulyErz/Facies-classification-from-well-logs. В качестве дальнейшего направления исследований можно, например, рассмотреть задачу восстановления пропущенных каротажных измерений с помощью различных глубоких нейронных сетей.

Список терминов

Каротаж (франц. carottage, от carotte — буровой кёрн, буквально — морковь) — геофизические исследования скважин, выполняемые с целью изучения геологических разрезов и выявления полезных ископаемых.

Каротажная диаграмма — визуализация результатов геофизических исследований в скважине, которая представляет собой кривые изменения различных физических параметров (или показаний скважинных приборов) вдоль разреза скважины.

Кёрн — образец горной породы, получаемый путем кольцевого разрушения забоя скважин при бурении.

Литология — важная часть петрографии, изучающая состав, структуру, происхождение и изменение осадочных пород; изучает закономерности и условия образования геологических осадков, процессы консолидации и литификации.

Наклономер (инклинометр, измеритель наклона) — средство измерения, предназначенное для контроля абсолютного или относительного углового положения объекта относительно вертикали.

Фация — слой или группа слоев, отражающих среду осадконакопления.

Список литературы

1. A. Al-Anazi, I.D. Gates, “A support vector machine algorithm to classify lithofacies and model permeability in heterogeneous reservoirs,” Engineering Geology 114 (3–4), pp. 267–277, 2010.
2. A. Bhatt, H.B. Helle, “Determination of facies from well logs using modular neural networks,” Petroleum Geoscience 8, pp. 217–228, 2002.
3. A. Hall, “Facies classification using machine learning,” The Leading Edge (Society of Exploration Geophysicists), 35(10), pp. 906–909, 2016.
4. A. Hall, M. Hall, “Distributed collaborative prediction: Results of the machine learning contest,” The Leading Edge (Society of Exploration Geophysicists), 36(3), pp. 267–269, 2017.
5. A. Hallam, “Facies Interpretation and the Stratigraphic Record,” Oxford, San Francisco: Freeman, 291 p, 1981.
6. A. Kakouei, M. Masihi, B.S. Sola, and E. Biniaz, “Lithological Facies Identification in Iranian Largest Gas Field: A Comparative Study of Neural Network Methods,” Journal Geological Society of India 84, pp. 326-334, 2014.
7. A. Natekin and A. Knoll, “Gradient boosting machines, a tutorial,” Front. Neurorobot., vol. 7, p. 21, 2013.
8. D. W. Marquardt, R. D Snee, “Ridge regression in practice,” American Statistician, 29, pp. 3-19, 1975.
9. GridSearchCV[Электронный ресурс] URL: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html (дата обращения: 14.03.2019).
10. H. Demuth, M. Beal, and M. Hagan, Neural network toolbox 6 (User’s Guide), The MathWorks, 2009.
11. I.K. Yeo and R.A. Johnson, “A new family of power transformations to improve normality or symmetry,” Biometrika, 87(4), pp. 954-959, 2000.
12. J. Jiajun, C.J. Scott, C.A. Stacy, “Geologic Facies Classification from Well Logs Using a Bi-directional Long Short-term Memory Neural Network,” unpublished, 2018. [Электронный ресурс] URL: https://www.researchgate.net/publication/327537571 (дата обращения: 18.12.2018)
13. J. Outrata, “Preprocessing Input Data for Machine Learning by FCA,” in *CLA*, pp. 187– 198, 2010.
14. J. R. Yeh, J. S. Shieh, N. E. Huang, “Complementary ensemble empirical mode decomposition: A novel noise enhanced data analysis method,” Advances in adaptive data analysis, T. 2. – №. 02. – pp. 135-156, 2010.
15. J. Rebeschini, M. Querales, G.A. Carvajal, M. Villamizar, F. Ma Aanan, J. Rodriguez, S. Knabe, F. Rivas, L. Saputelli, A. Al-Jasmi, H. Nasr, H.K. Goel,

"Building neural-network-based models using nodal and time-series analysis for short-term production forecasting," Society of Petroleum Engineers - SPE Intelligent Energy International 2013: Realising the Full Asset Value, pp. 102-114, 2013.
16. J. Schmidhuber, "Deep learning in neural networks: An overview," Neural Networks, 61: pp. 85–117, 2015.
17. J.L. Baldwin, R.M. Bateman, C.L. Wheatley, "Application of a neural network to the problem of mineral identification from well logs," The Log Analyst 31(5), pp. 279–293, 1990.
18. J.M. Busch, W.G. Fortney, L.N. Berry, "Determination of lithology from well logs by statistical analysis," SPE Formation Evaluation 2, pp. 412–418, 1987.
19. K.P. Dorrington, C.A. Link, "Genetic-algorithm/neural network approach to seismic attribute selection for well-log prediction," Geophysics 69, pp. 212–221, 2004.
20. KNeighborsRegressor [Электронный ресурс] URL: https://scikit-learn.org/0.20/modules/generated/sklearn.neighbors.KNeighborsRegressor.html#sklearn.neighbors.KNeighborsRegressor (дата обращения: 07.03.2019).
21. L. Bremen, "Random Forests," Machine Learning, 45(1), pp. 5-32, 2001.
22. L. E. Raileanu and K. Stoffel, "Theoretical comparison between the gini index and information gain criteria," Univeristy of Neuchatel, 2000.
23. M. Wolf, J. Pelissier-Combescure, "Faciolog-automatic electrofacies determination," Transactions of the SPWLA Annual Logging Symposium (Society of Professional Well Log Analysts) 2, 23 p., 1982.
24. M.A. Colominas, G. Schlotthauer, M.E. Torres, P. Flandrin, "Noise-assisted emd methods in action," Adv Adapt Data Anal; 04(04), 2012.
25. M.J. Pyrcz, C.V. Deutsch, Geostatistical Reservoir Modeling. Oxford University Press, 2014.
26. M.K. Dubois, A.P. Byrnes, G.C. Bohling, J.H. Doveton, "Multiscale geologic and petrophysical modeling of the giant Hugoton gas field (Permian), Kansas and Oklahoma, USA," AAPG Memoir 88/SEPM Special Publication, pp. 307–353, 2006.
27. M.K. Dubois, G.C. Bohling, S. Chakrabarti, "Comparison of four approaches to a rock facies classification problem," Computers & Geosciences 33, pp. 599–617, 2007.
28. M.M. Saggaf, and E.L. Nebrija, "Estimation of lithologies and depositional facies from wire-line logs," AAPG Bulletin (American Association of Petroleum Geologists), 84(10), pp. 1633–1646, 2000.
29. N. Bhatia et al, "Survey of Nearest Neighbor Techniques," International Journal of Computer Science and Information Security, Vol. 8, No. 2, 2010.

30. N.E. Huang, Z. Shen, S.R. Long, M.C. Wu, H.H. Shih, Q. Zheng, N-C. Yen, C.C. Tung, H.H. Liu, “The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis,” Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences, T. 454. – №. 1971. – pp. 903-995, 1998.
31. P. Delfiner, O. Peyret, O. Serra, “Automatic determination of lithology from well logs,” SPE Formation Evaluation 2, pp. 303–310, 1987.
32. P. Wong, F. Jian, I. Taggart, “A critical comparison of neural networks and discriminant analysis in lithofacies, porosity and permeability predictions,” Journal of Petroleum Geology 18, pp. 191–206, 1995
33. R. Tibshirani, “Regression shrinkage and selection via the lasso,” Journal of the Royal Statistical Society. Series B. Methodological, 58(1): pp. 267–288, 1996.
34. R.E. Schapire, “Explaining adaboost,” *Empirical inference*. Springer, Berlin, Heidelberg, pp. 37-52, 2013.
35. S. B. Kotsiantis, D. Kanellopoulos, and P. E. Pintelas, “Data preprocessing for supervised leaning,” International Journal of Computer Science 1.2, pp. 111– 117, 2006.
36. S. Gaci, “A new ensemble empirical mode decomposition (EEMD) denoising method for seismic signals,” *Energy Procedia, 97,* pp. 84– 91, 2016.
37. S. S. Shapiro, M. B. Wilk, “An analysis of variance test for normality,” Biometrika, 52, №3 — pp. 591-611, 1965.
38. S.J. Rogers, J. Fang, C. Karr, D. Stanley, “Determination of lithology from well logs using a neural network (1),” AAPG Bulletin 76, pp. 731–739, 1992.
39. T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” In Proceedings of the 22nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining, pp. 785–794, San Francisco, CA, 2016.
40. T. Darling, Well Logging and Formation Evaluation. Gulf Professional Publishing, 2005, 336 p.
41. T. Liu, S. Fang,Y. Zhao, P. Wang, J. Zhang, “Implementation of training convolutional neural networks,” arXiv preprint arXiv:1506.01195, 2015.
42. T. M. Cover and P. E. Hart, “Nearest Neighbor Pattern Classification,” IEEE Trans. Inform. Theory, Vol. IT-13, pp. 21-27, 1967.
43. T. Young, D. Hazarika, S. Poria, E. Cambria, “Recent Trends in Deep Learning Based Natural Language Processing,” arXiv preprint arXiv:1708.02709, 2017.
44. Total energy consumption. Acceleration in energy consumption in 2017. [Электронный ресурс] URL: https://yearbook.enerdata.net/total-energy/world-consumption-statistics.html (дата обращения: 05.11.2018)

45. V.A Davydov, A.V. Davydov, “Management of Empirical Mode Decomposition of signals in the analysis and processing of geophysical data,” Karotazhnik, no. 5, pp. 98–114, 2010. In Rus.
46. Y. Imamverdiyev, L. Sukhostat, “Lithological facies classification using deep convolutional neural network,” Journal of Petroleum Science and Engineering 174, pp. 216 – 228, 2019.
47. Y. LeCun, L. Bottou, P. Haffner, “Gradient-based learning applied to document recognition,” Proceedings of the IEEE, 86(11), pp. 2278–2324, 1998.
48. Y. Li, R.A. Sprecher, “Facies identification from well logs: A comparison of discriminant analysis and naïve Bayes classifier,” Journal of Petroleum Science and Engineering 53, pp. 149-157, 2006.
49. Z. Huang, M.A. Williamson, “Artificial neural network modeling as an aid to source rock characterization,” Marine and Petroleum Geology 13, pp. 227-290, 1996.
50. Z. Wu, N.E. Huang, “Ensemble empirical mode decomposition: a noise-assisted data analysis method,” Adv Adapt Data Anal; 01(01): 1–41, 2009.
51. А.С. Долгаль, Л.А. Христенко, “Применение эмпирической модовой декомпозиции при обработке геофизических данных,” Известия Томского политехнического университета. Инжиниринг георесурсов, Т. 328. № 1. 100–108, 2017.
52. Ф.А. Мохаммед, “Очистка сигналов датчиков от шумов при проведении физических экспериментов методом эмпирической модовой декомпозиции,” диссертация на соискание степени магистра, 2016.