Федеральное государственное автономное образовательное учреждение
высшего образования «Национальный исследовательский университет
«Высшая школа экономики»

Факультет компьютерных наук
Основная образовательная программа
Прикладная математика и информатика

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

Программный проект на тему

"Машинное обучение и Рамановская спектроскопия в диагностике опухолевых тканей"

Выполнил студент группы 198, 4 курса,
Гладких Роман Евгеньевич

Руководитель ВКР от НИУ ВШЭ:
Доцент ФКН НИУ ВШЭ
Соколов Евгений Андреевич

Соруководитель ВКР:
Начальник отдела ИТ ИПЛИТ РАН
Воронин Игорь Вадимович

Москва 2023

1 Аннотация

1.1 На русском языке

В настоящее время машинное обучение является одной из самых быстрорастущих областей науки, и оно уже продемонстрировало свой большой потенциал для использования в медицине. Рамановская спектроскопия — это спектроскопический метод исследования, используемый для определения колебательных состояний внутри образцов, однако, имеется мало сообщений о его применении в живых тканях. Вследствие чего на данный момент Рамановская спектроскопия является одним из наиболее перспективных методов диагностики опухолевых тканей на ранней стадии болезни, путем выявления характерных изменений в составе ткани при раке.

В данной работе описывается процесс создания и исследования моделей для диагностики опухолевых тканей на базе алгоритмов классификации машинного обучения, основанных на моделях логистической регрессии, бустинга и случайного леса. Также в данной работе при помощи алгоритмов отбора признаков выявляются наиболее важные Рамановские сдвиги, с целью улучшения интерпретируемости модели и создания механизма поддержки врачебных решений.

В качестве места хранения и работы с кодом был выбран сервис GitHub - https://github.com/RomanGladkikh/ML-and-RS-in-the-diagnosis-of-cancer.

Ключевые слова — Рамановская спектроскопия; нейроонкология; классификация; машинное обучение

1.2 In English

Currently, machine learning is one of the fastest growing fields of science, and it has already demonstrated its great potential for use in medicine. Raman spectroscopy is a spectroscopic research method used to determine vibrational states inside samples, however, there are few reports of its use in living tissues. As a result, at the moment, Raman spectroscopy is one of the most promising methods for diagnosing tumor tissues at an early stage of the disease, by identifying characteristic changes in the composition of tissue in cancer.

This paper describes the process of creating and researching models for the diagnosis of tumor tissues based on machine learning classification algorithms based on logistic regression, boosting and random forest models. Also, in this work, the most important Raman shifts are identified using feature selection algorithms in order to improve the interpretability of the model and create a mechanism to support medical decisions.

The GitHub service was chosen as a place to store and work with the code - https://github.com/RomanGladkikh/ML-and-RS-in-the-diagnosis-of-cancer.

Keywords — Raman spectroscopy; neuro-oncology; classification; machine learning

Содержание

1 Аннотация **1**

1.1 На русском языке 1

1.2 In English 2

2 Терминология **4**

3 Введение **5**

4 Обзор литературы **7**

4.1 Обзор предметной области 7

4.2 Обзор литературы 8

4.3 Обзор современных подходов 9

5 Постановка задачи **12**

6 Анализ данных **13**

6.1 Среднее и медиана 14

6.2 Рассмотрение различных перцентилей классов 16

7 Проведение экспериментов **18**

7.1 Использование ML методов 18

7.2 Использование алгоритмов кластеризации 20

8 Отбор признаков **21**

8.1 Взаимная информация 22

8.2 Exhaustive Feature Selector 24

8.3 Recursive Feature Elimination 25

8.4 Random Forest Importance 26

8.5 Principal Component Analysis 29

8.6 Ансамбль методов 30

9 Заключение **31**

2 Терминология

1 Рамановская спектроскопия (Raman spectroscopy) - это спектроскопический метод исследования, используемый для определения колебательных состояний внутри образцов.

2 Рамановский сдвиг (Raman Shift) - разница между энергией падающего фотона и энергией рассеянного фотона.

3 ML (Machine Learning) - термин для обозначения методов, относящихся к машинному обучению и искусственному интеллекту.

4 DL (Deep learning) - является частью более широкого семейства методов машинного обучения, которое основано на искусственных нейронных сетях.

5 Бейзлайн (от англ. baseline) - термин для обозначения качества референсного решения некоторой задачи.

6 Метрика (Metric) — это качественный или количественный показатель, который отражает ту или иную характеристику и уровень успешности алгоритма или модели.

3 Введение

Рак — это большая группа заболеваний, которые могут начаться практически в любом органе или ткани организма, когда больные клетки бесконтрольно растут, выходят за свои обычные границы, проникают в соседние части тела и/или распространяются на другие органы. Это также вторая по значимости причина смертности во всем мире, на долю которой приходится, по оценкам, 10 миллионов смертей, или каждая шестая смерть[1]. Частота диагностики рака растет с каждым годом. Это связано как с совершенствованием методов диагностики, так и с влиянием факторов, провоцирующих такие заболевания.

В 2021 году 3,23 миллиона жителей России находились под наблюдением врачей. В том же году российские врачи поставили 580 тысяч диагнозов, связанных со злокачественными новообразованиями. Каждый год в нашей стране 300 тысяч человек умирают от рака. Причинами высокой смертности являются поздняя диагностика опухолей из-за того, что пациент не обращает внимания на симптомы заболевания или откладывает обращение к врачу.

Сегодня множество больших компаний, корпораций и даже целых стран вкладывают ресурсы в развитие методов борьбы с раком. В рамках этих работ учеными исследуются и тестируются различные гипотезы, направленные как на улучшение методов борьбы с раком на поздних стадиях, так и на повышение точности ранней диагностики, за счёт поиска новых методов и развития уже существующих. Ранняя диагностика рака имеет решающее значение для эффективного и успешного лечения, поскольку более 90% всех пациентов излечиваются при получении своевременного лечения. Рамановская спектроскопия (RS) является одним из ведущих методов для ранней диагностики, так как изменения в структуре и концентрации биохимических веществ клеток начинаются гораздо раньше, чем появляются первые клинические признаки злокачественной опухоли. Таким образом, RS позволяет обнаружить измене-

[1] https://www.who.int/health-topics/cancer

ния, вызванные раком в тканях на ранней стадии болезни. К преимуществу RS относится тот факт, что это неинвазивный метод диагностики.

Однако, на данный момент существуют значительные пробелы в применении RS, поскольку до сих пор не существует полного сопоставления всех Рамановских сдвигов и конкретных функциональных элементов, с которыми они связаны. Большинство ныне существующих исследований проведены в диапазоне [100, 2000], таким образом, исследование Рамановских сдвигов в используемом нами диапазоне [2000, 3600], может повлечь за собой открытие новых закономерностей в данных, из-за недостаточного количества работы, проведенной в данном направлении.

В рамках данной работы будет проведен анализ различных алгоритмов Машинного обучения (Logistic regression, SVM, Random forest, CatBoost, K-means), на задаче классификации здоровых и опухолевых тканей. Дополнительно будет проведен анализ и поиск новых и наиболее важных Рамановских сдвигов с помощью различных методов машинного обучения (Взаимная информация, Exhaustive Feature Selector, Recursive Feature Elimination, Random Forest Importance, Principal Component Analysis).

4 Обзор литературы

4.1 Обзор предметной области

Как правило, задачи по развитию методов по борьбе с раком решаются внутри частных и государственных компаний, вследствие чего исследования обычно включают только количественное измерения этих результатов, в отличие от публикации программных решений, которые зачастую не являются частью статей о результатах исследований в связи с политикой неразглашения этих компаний. В области использования машинного обучения и Рамановской спектроскопии в диагностике опухолевых тканей проведено не так много исследований. Тем не менее для данной задачи успешно применены основные алгоритмы машинного обучения, а также предложены некоторые подходы для использования нейронных сетей для классификации тканей. Из таких публикаций мы рассмотрим интуицию методов и идей, и будем развивать ее в рамках целеполагания текущей задачи. Тем не менее, так как задачи бинарной классификации тканей сводятся к применению традиционных методов машинного обучения, в качестве методологической литературы будем рассматривать современные статьи по машинному обучению, применимые к нашим задачам. Для диагностики опухолевых тканей с помощью Рамановской спектроскопии можно выявить несколько основных базовых задач:

1 Задача классификации тканей на здоровые и опухолевые.

2 Задача интерпретации основных факторов предсказательных моделей для выявления ключевых Рамановских сдвигов и последующего их исследования.

4.2 Обзор литературы

Одна из работ, на которую опиралось наше исследование - Riva et al. (2021), в которой группа итальянских ученых во главе с М. Ривой работала с 3450 образцами биопсии глиомы в спектральном диапазоне [90, 1800] cm^{-1}. Используя алгоритмы машинного обучения: Random Forest и Gradient boosting trees, были обнаружены 19 важных новых Рамановских сдвига в диапазоне [430, 1603] cm^{-1}. Точность окончательной модели составила 83%.

Также в работах Huang et al. (2003) и Huang et al. (2005) была предоставлена информацию для интерпретации Рамановских сдвигов. В зависимости от типа ткани эти пики могут отличаться по широте и высоте, поскольку типы опухолей различаются по количеству таких веществ, как белки, липиды и нуклеиновые кислоты.

Так в работе Gniadecka et al. (2004) было проведено исследование меланомы с использованием RS и нейронных сетей. Анализ чувствительности RS, используемый нейронной сетью, был исследован для определения важности отдельных компонентов в RS. Было обнаружено, что интенсивность Рамановского сдвига белка амида I уменьшилась примерно на 1660 cm^{-1}. Меланома также показала увеличение интенсивности примерно на 1310 cm^{-1}, а базальноклеточный рак - примерно на 1330 cm^{-1} соответственно, что указывает на гипотетический метод отделения больных клеток от здоровых.

Исследование Jermyn et al. (2016) рака головного мозга проводилось в присутствии артефактов освещения операционной с использованием RS. Были обнаружены важные различия между здоровыми образцами и образцами, пораженными раком. Эти различия были выражены в несоответствиях в белках (1005 cm^{-1}), нуклеиновой кислоте ([1540, 1645] cm^{-1}), холестерине (700 cm^{-1}) и фосфолипидах (1142 cm^{-1}).

Кроме того, исследование рака молочной железы проводилось с использованием RS и 2D сверточной нейронной сети Shang et al. (2022). В результате этого исследования были получены новые пики для Рамановских сдвигов в

диапазоне [1247, 1318] cm^{-1}, а также 1660 cm^{-1}. В результате для выборки была достигнута средняя точность распознавания 96,01%.

Таким образом, очень важно провести анализ Рамановских сдвигов, так как во всех вышеприведённых исследованиях их пики находились в диапазоне [430, 1660] cm^{-1}, в то время как наш диапазон исследования [2000, 3600] cm^{-1}, вследствие чего мы можем обнаружить новые Рамановские сдвиги. Также в дальнейшем при помощи медицинских лабораторий можно сопоставить пики Рамановских сдвигов и функциональные элементы, которым они соответствуют. Благодаря этому мы сможем точнее интерпретировать пики и получать более надежные результаты.

4.3 Обзор современных подходов

Бинарная классификация тканей с помощью Рамановской спектроскопии - очень сложная и важная задача. Основное внимание уделяется обнаружению таких Рамановских сдвигов, где различия между пораженной и здоровой тканью будут показательными. В настоящий момент времени не существует подхода, однозначно признанного лучшим для решения данной задачи. Лучший подход определяется путем проведения ряда экспериментов над моделями. Рассмотрим несколько моделей машинного обучения, являющихся основополагающими в современной науке

1 Logistic regression (LR) - одна из простейших моделей бинарной классификации. В данной модели целевая переменная принимает только два значения “0” или “1”, в нашем случае здоровая или больная ткань соответственно. LR быстро обучается и легко интерпретируема, но основной минус LR заключается в опоре на предположении о линейной зависимости между целевой переменной и признаками, а также проблемами с переобучением в случае большего количества признаков и меньшего количества наблюдений.

2 Support vector machine (SVM) - модель, принадлежащая семейству ли-

нейных классификаторов. Работа данной модели основывается на поиске с помощью опорных векторов, такой гиперплоскости, что расстояние между опорными векторами будет максимизировано. SVM устойчив к выбросам, так как только опорные вектора определяют гиперплоскость, также он показывает хорошие результаты при большом количестве признаков. В исследовании Widjaja et al. (2008) при классификации тканей толстой кишки с использованием RS и SVM была получена точность 99%.

3 Random forest (RF) - ML алгоритм, основанный на использовании большого ансамбля решающих деревьев, каждое из которых по отдельности выдает слабое качество при классификации, но при объединении большого количества таких решающих деревьев мы получаем хорошие результаты. Эффективность этого алгоритма в применении к классификации раковых клеток с использованием RS была подтверждена в исследованиях Riva et al. (2021) и Seifert (2020).

4 CatBoost - алгоритм, основанный на градиентном бустинге (GB). GB используется для уменьшения смещения и дисперсии, мы последовательно обучаем алгоритмы каждый из которых уменьшает ошибки предыдущих, таким образом, мы преобразуем много слабых алгоритмов в один сильный ансамбль. CatBoost использует oblivious деревья решений, чтобы вырастить сбалансированное дерево. В исследовании Riva et al. (2021) указана эффективность применения GB к диагностике биопсии глиомы, несмотря на то что формально GB слабее нейронных сетей, он все же выигрывает в удобстве применения, размере и интерпретируемости модели.

5 Алгоритм K-Means (K-Средних) - один из наиболее простых методов кластеризации в классической реализации. Он разбивает множество элементов векторного пространства на заранее известное число кластеров k. Алгоритм стремится минимизировать среднеквадратичное откло-

нение на точках каждого кластера. Основная идея заключается в том, что на каждой итерации перевычисляется центр масс для каждого кластера, полученного на предыдущем шаге, затем векторы разбиваются на кластеры вновь в соответствии с тем, какой из новых центров оказался ближе по выбранной метрике. В пространствах с очень высокой размерностью евклидовы расстояния имеют тенденцию к завышению (это пример так называемого “проклятия размерности”). Запуск алгоритма уменьшения размерности, такого как анализ главных компонент (PCA), перед кластеризацией k-средних может облегчить эту проблему и ускорить вычисления.

6 Artificial neural networks (ANN) - алгоритм, который имитирует строение нервной ткани. ANN состоит из множества “нейронов”, которые могут взаимодействовать друг с другом, при этом поток информации в ANN движется только в одну сторону. Таким образом, “нейрон” получает сигнал в виде вещественного числа, после чего обрабатывает его. Выходной сигнал каждого нейрона представляет собой некоторую нелинейную функцию от входных сигналов. Нейроны и ребра между ними имеют вес, который меняется в процессе обучения. После каждой итерации ANN подсчитывает точность, затем с помощью функции потерь меняет веса, после чего продолжает работу Seifert (2020).

7 Convolutional neural network (CNN) - данная нейронная сеть, содержит в себе хотя бы один сверточный слой. Такие слои могут быть либо полностью связаны, либо объединены. При этом CNN считается сильнее, чем ANN, а также не имеет ограничения в виде однонаправленности, что разрешает петли и циклы. С помощью CNN и RS было проведено исследование рака молочной железы Shang et al. (2022), в результате для выборки была достигнута средняя точность распознавания 96,01%. Также в исследовании Lee et al. (2019) при определении рака предстательной железы была получена точность = 93%.

5 Постановка задачи

Исследования проводились на 1022 наблюдениях, с соотношением здоровых тканей к больным, как $432 : 590$ $(1 : 1.36)$. Каждому наблюдению соответствует Рамановская спектроскопия, состоящая из 1015 признаков-пар: длина волны - интенсивность.

В этой работе будет рассмотрено множество алгоритмов классификации тканей на основе данных по Рамановской спектроскопии с применением методов классического машинного обучения и выбран лучший. Также будет произведен поиск наиболее важных Рамановских сдвигов при помощи методов отбора признаков.

В нашей задаче целевой показатель - индикатор отношения ткани к здоровым или опухолевым.

6 Анализ данных

Формально мы имеем 1022 графика, соответствующих наблюдениям, по 1015 признаков. Иллюстрация исходного набора данных приведена на рисунке 6.1. Всего предоставлено 432 наблюдения, соответствующих здоровым тканям и 590 наблюдений, соответствующих опухолевым.

	2141	2142	2144	2145	2147	2149	2150	2152	2153	2155	...	3540	3541	3542	3544	3545	3546	3547	3549	3550	target
0	30525.419922	30643.562500	30689.888672	30643.578125	30683.531250	30421.400391	30443.724609	30731.535156	30566.783203	30630.755859	...	32005.250000	32313.267578	32242.830078	32017.634766	32374.712891	31937.625000	31889.439453	32100.001953	32236.521484	0
1	37439.214844	37367.546875	37552.988281	37626.691406	37489.488281	37281.878906	37405.121094	37202.175781	37061.503906	37531.996094	...	38717.175781	38918.777344	38739.839844	38241.191406	38843.953125	38899.093750	38944.066406	38799.558594	38987.015625	0
2	39250.558594	39315.093750	39397.226563	39287.679688	39467.324219	39576.726563	39583.847656	39296.699219	39377.335938	39320.394531	...	36343.796875	36467.566406	36788.820313	36637.882813	36635.527344	36370.425781	36673.582031	36084.382813	36416.519531	0
3	33261.566406	33528.921875	33580.792969	33556.003906	33510.425781	33456.828125	33217.730469	33335.144531	33650.976563	33496.539063	...	37175.000000	37653.027344	37192.933594	37215.128906	36954.292969	37328.792969	37577.144531	37674.816406	37436.417969	0
4	33028.371094	33128.003906	32876.304688	33137.304688	33061.253906	33045.902344	33000.160156	33024.757813	32934.187500	33051.582031	...	30153.716797	30371.800781	30313.160156	30291.136719	30391.275391	30025.048828	30318.695313	30317.015625	30292.564453	0
5	19338.117188	19196.261719	19346.613281	19404.382813	19463.775391	19302.617188	19354.017578	19186.345703	19428.099609	19394.763672	...	17214.128906	17529.158203	17370.074219	17200.740234	17065.949219	17077.740234	17250.466797	17241.923828	17406.966797	0
6	15217.953125	15142.085938	15203.535156	15271.405273	15085.225586	15001.247070	14854.868164	14805.966797	14840.221680	15036.102539	...	5137.179199	5115.893066	5015.216797	4979.643555	5045.868164	4871.789063	4986.909180	5028.724609	4957.409668	0
7	35759.816406	36075.726563	35749.734375	36092.957031	35824.316406	35760.132813	35953.328125	35684.433594	36114.457031	36103.070313	...	27761.187500	28239.236328	27956.246094	27748.457031	27721.802734	27691.062500	27782.535156	27759.937500	27869.966797	0
8	25421.126953	25410.195313	25346.546875	25434.638672	25383.744141	25152.189453	25510.507813	25240.478516	25237.462891	25096.828125	...	18203.523438	18246.080078	18258.123047	17958.675781	18184.472656	18159.839844	18067.968750	18112.578125	18196.121094	0
9	34051.101563	34136.632813	34509.707031	34135.277344	34268.832031	34483.972656	34157.199219	34135.718750	34470.964844	34774.378906	...	42349.894531	41774.898438	42268.265625	42208.078125	42074.550781	42146.675781	41937.695313	42095.390625	42254.328125	0
10	34139.558594	34319.320313	33992.867188	34219.007813	34040.964844	33864.425781	33690.992188	34097.996094	33974.113281	34194.183594	...	33797.671875	33943.277344	33865.046875	33970.003906	34235.867188	33876.570313	33765.648438	34080.527344	33950.125000	0
11	84098.601563	83815.554688	83701.656250	83856.093750	84068.140625	83816.726563	83774.632813	83778.890625	83751.140625	83966.601563	...	80319.773438	79881.101563	79857.468750	79901.000000	79791.812500	79721.273438	80023.500000	79622.859375	80036.507813	0
12	79249.234375	79311.429688	79001.507813	79247.382813	79113.046875	78400.000000	78799.375000	78570.445313	78910.867188	78920.242188	...	46332.433594	46492.929688	46246.406250	46026.238281	46510.523438	46249.574219	46004.804688	46234.679688	46238.757813	0
13	41897.382813	42222.886719	42182.792969	42216.164063	42335.843750	41940.875000	42309.964844	41890.785156	42079.277344	41936.636719	...	30547.265625	30130.511719	30214.300781	30294.048828	30447.134766	30368.093750	30362.353516	30271.027344	30393.285156	0
14	16062.418945	16066.610352	16059.622070	16143.691406	15909.799805	15993.871094	16023.624023	16098.169922	16247.938477	16074.564453	...	16304.695313	16332.601563	16256.711914	16160.429688	16117.047852	16189.738281	16187.083008	16272.040039	15996.967773	0
15	35001.449219	34585.847656	35019.718750	34969.746094	34902.164063	34743.429688	34544.644531	34600.031250	34660.242188	34590.875000	...	30536.208984	30461.339844	30445.460938	30590.404297	30664.134766	30597.361328	30506.111328	30441.306641	30649.138672	0
16	32963.957031	33044.414063	33137.691406	33542.550781	33207.718750	33277.121094	33309.785156	33147.343750	33155.988281	33519.867188	...	41475.023438	41319.054688	41295.320313	41497.554688	41121.550781	41160.886719	41157.449219	40880.964844	41054.164063	0
17	79700.507813	79603.132813	79572.796875	80371.609375	80424.539063	80140.203125	79766.179688	79836.554688	79687.804688	79988.609375	...	74879.265625	74984.820313	74697.562500	74605.617188	74837.406250	74525.468750	74529.156250	74667.320313	75091.148438	0
18	49397.476563	49384.820313	49416.871094	49162.964844	49377.140625	49658.554688	49011.183594	49148.707031	49163.121094	49335.953125	...	38766.003906	38662.371094	38625.855469	38906.914063	38501.859375	38719.863281	38508.117188	38455.207031	38601.914063	0
19	15238.749023	15260.290039	15221.147461	15314.583984	15187.541992	15260.236328	15332.966797	15175.439453	15195.393555	15354.565430	...	17407.455078	17476.294922	17374.191406	17500.052734	17273.693359	17544.191406	17486.822266	17445.726563	17588.357422	0
20	40031.687500	39989.441406	39931.195313	40085.398438	39918.472656	39980.035156	40027.238281	39697.031250	39652.996094	39994.554688	...	39986.726563	39392.925781	39808.425781	39739.570313	39835.621094	39840.050781	40139.871094	39694.054688	39541.238281	0
21	37098.816406	37087.738281	36790.746094	37174.195313	37065.632813	36507.839844	36701.234375	36691.613281	37067.523438	36653.148438	...	17920.714844	18431.267578	18227.589844	18148.013672	18046.001953	17935.794922	18249.279297	18031.072266	18344.867188	0
22	18357.962891	18209.462891	18318.031250	18457.007813	18460.224609	18556.150391	18199.652344	18322.744141	18231.541016	18361.101563	...	14414.620117	14370.898438	14589.722656	14596.911133	14422.875000	14586.854492	14543.114258	14354.685547	14558.443359	0
23	28632.037109	28858.263672	28758.783203	28745.515625	29045.451172	28808.468750	28648.103516	28409.287109	28811.816406	28700.921875	...	27417.279297	27386.511719	27158.294922	27269.962891	27481.451172	27104.419922	27220.210938	27152.667969	27456.265625	0
24	10516.717773	10500.300781	10474.291992	10465.840820	10433.415039	10543.231445	10466.039063	10398.400391	10559.480469	10461.427734	...	10435.086914	10482.086914	10647.159180	10531.356445	10464.373047	10647.874023	10595.144531	10552.579102	10558.691406	0

Рис. 6.1: Пример устройства данных

На рисунках 6.2 и 6.3 приведен пример графиков Рамановской спектроскопии для здоровой и опухолевой тканей. Общий вид графиков достаточно похож и пики обоих графиков, находятся примерно в одном интервале. Заметим также, что для данных примеров график здоровой ткани более гладкий, в отличие от графика опухолевой ткани, который сильно колеблется.

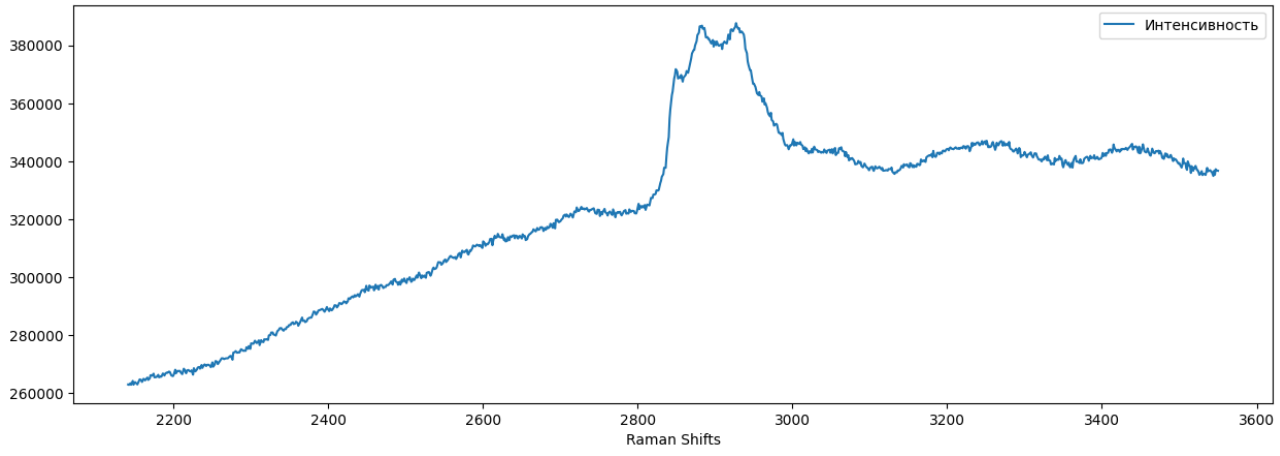

Рис. 6.2: Пример графика спектроскопии для здоровой ткани

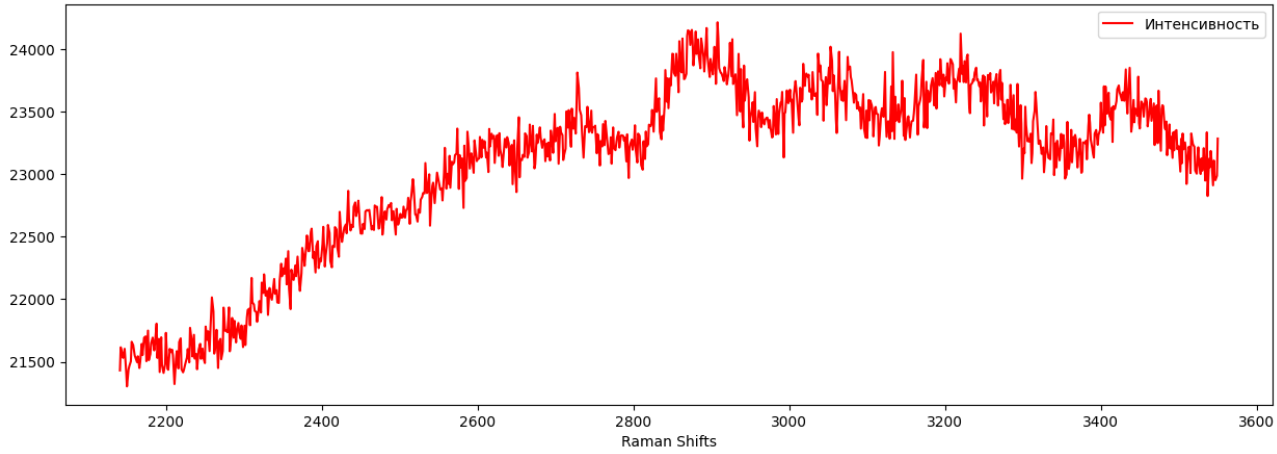

Рис. 6.3: Пример графика спектроскопии для больной ткани

В этом разделе мы сфокусируемся на сравнении графиков Рамановской спектроскопии для здоровых и больных тканей, с целью поиска закономерностей в данных и оценивания визуальной отделимости классов. Рассмотрим следующие аспекты:

1 Среднее и медиана классов по каждому признаку.

2 Рассмотрение различных перцентилей классов, во избежание выбросов в данных по признакам.

6.1 Среднее и медиана

На графике 6.4 изображено сравнение усредненных графиков для здоровых и больных тканей. Видно, что данные графики обладают следующими свойствами:

1 Усредненный график для здоровых тканей убывает кроме одного интервала, где наблюдается резкий рост, а потом падение до изначального тренда.

2 Усредненный график для больных тканей, в отличие от графика здоровых тканей возрастает. На интервале совпадающим с интервалом для графика здоровых тканей наблюдается резкий рост, а потом падение до изначального тренда.

3 С увеличением Raman Shift графики стремятся к пересечению друг с другом

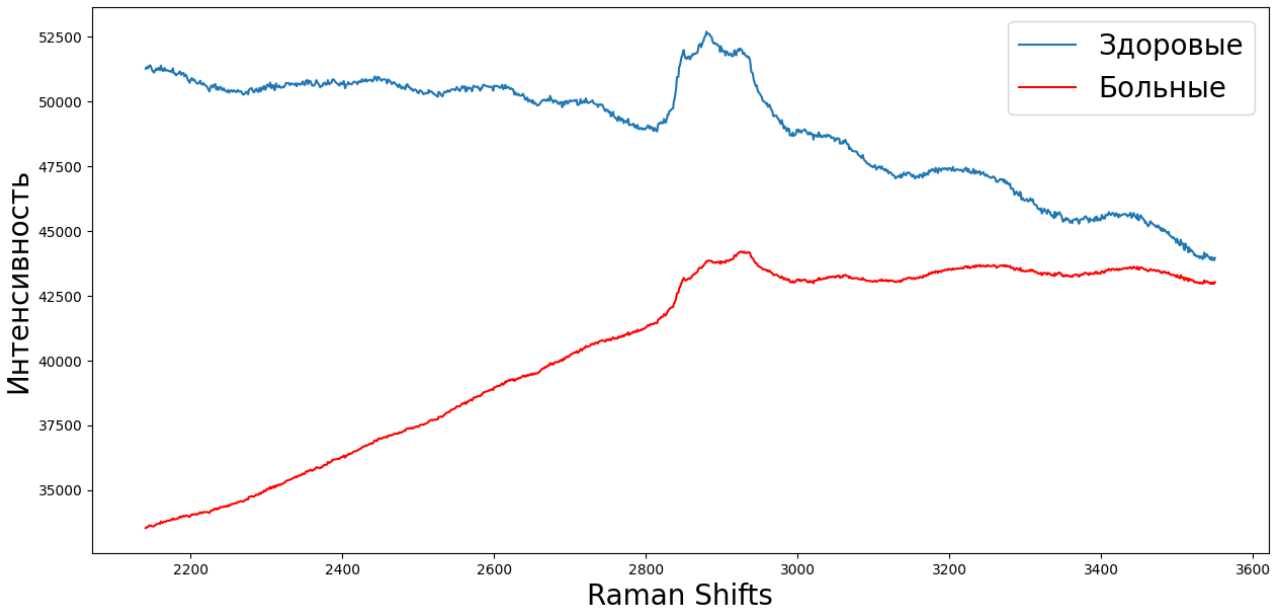

Рис. 6.4: Сравнение усредненных графиков Рамановской спектроскопии для здоровых и больных тканей

На графике 6.5 изображено сравнение медиан графиков для здоровых и больных тканей. Заметим, что данные графики обладают следующими свойствами:

1 С уменьшением Raman Shift становится проще различить случайные больные и здоровые ткани, вследствие этого самые важные признаки скорее всего будут находиться в интервале [2150, 2400] cm^{-1}, так как после 2400 cm^{-1} графики начинают быстро сближаться.

2 Аналогично, наименее информативные признаки будут находиться в интервале [3100, 3300] cm^{-1}, так как около 3200 cm^{-1} графики пересекаются.

В отличие от 6.4, на графике 6.5 можно заметить что данные ряды нельзя признать легкими для разделения. Для более точного анализа попробуем рассмотреть различные перцентили для распределения данных.

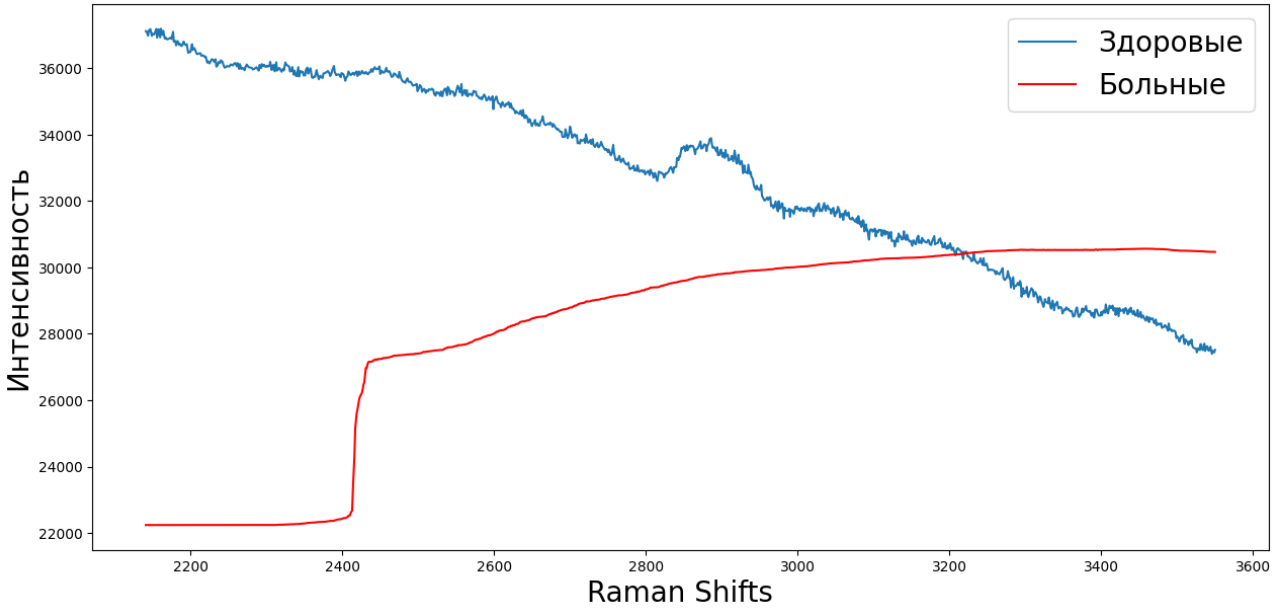

Рис. 6.5: Сравнение медиан графиков Рамановской спектроскопии для здоровых и больных тканей

6.2 Рассмотрение различных перцентилей классов

Рассмотрим 10% и 90% перцентили для классов во избежание выбросов в данных по признакам и получения более репрезентативной статистики. На графиках 6.6 и 6.7 указано сравнение средних, медианы, а также 10% и 90% перцентилей для здоровых и больных тканей соответственно.

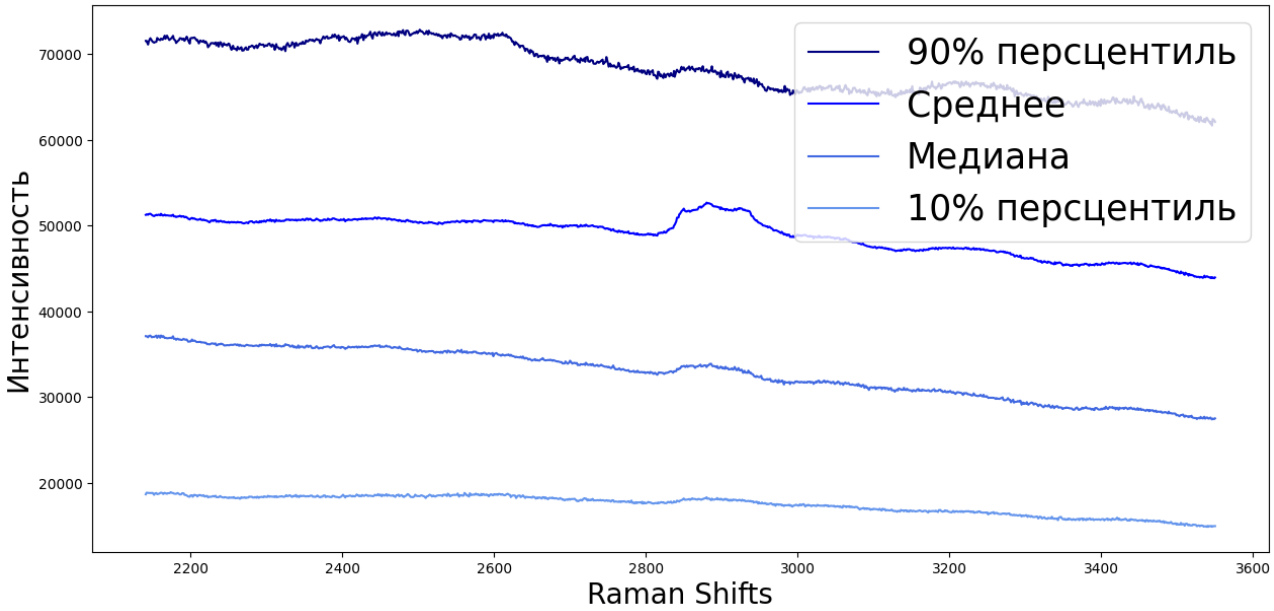

Рис. 6.6: Различные статистики для здоровых тканей

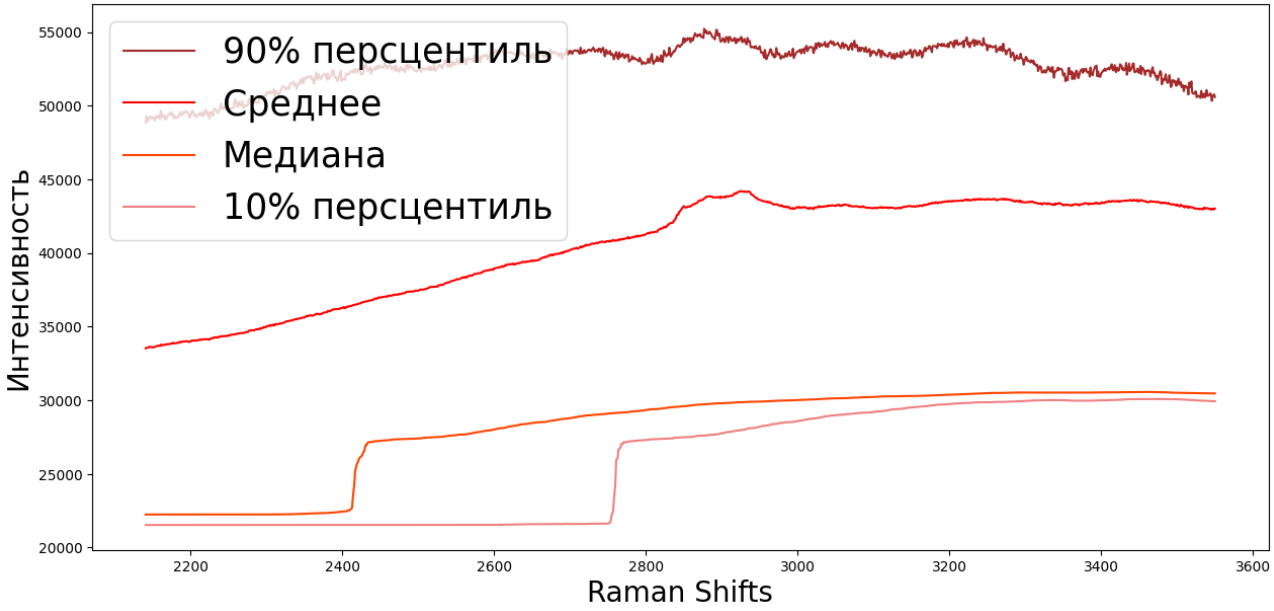

Рис. 6.7: Различные статистики для больных тканей

На графике 6.8 видно, что больные ткани распределены плотнее, чем здоровые, так как и 10% и 90% перцентили для больных тканей находятся между графиками функций 10% и 90% перцентилей здоровых тканей, что может усложнить разделение классов.

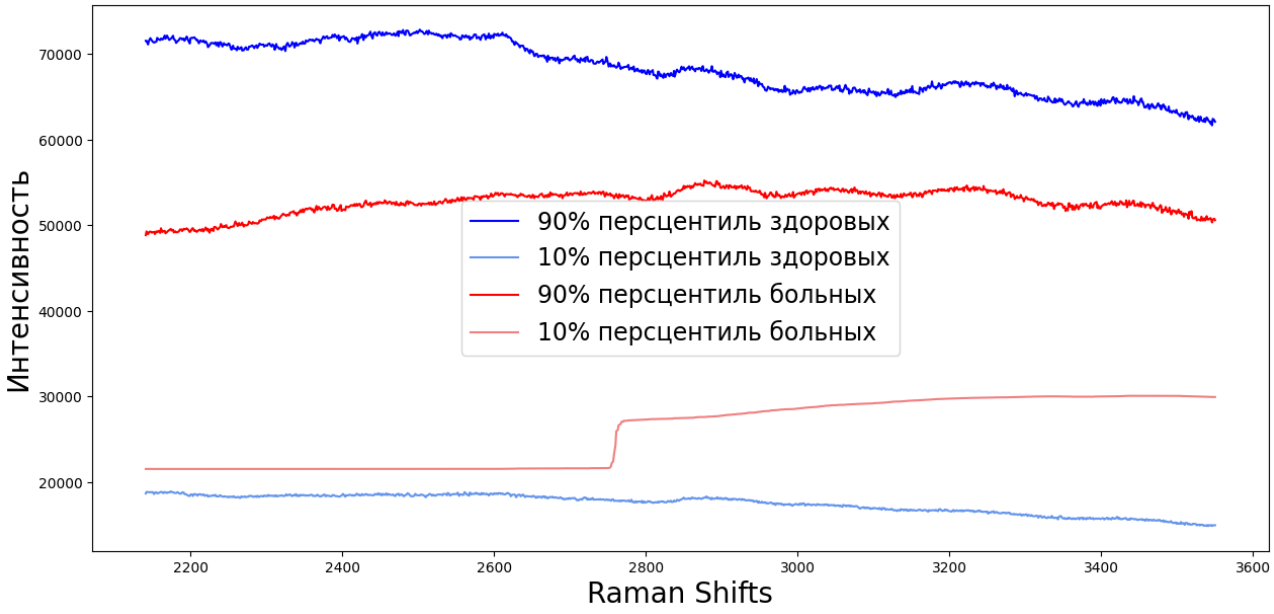

Рис. 6.8: Сравнение 10% и 90% перцентилей для классов

7 Проведение экспериментов

Для нашей задачи в силу медицинской специфики не существует единственной однозначно верной метрики качества, вследствие этого мы будем использовать несколько метрик для прогнозирования результата:

1 Accuracy (Точность) - доля правильных ответов алгоритма. Данная метрика интуитивно понятна, а также подходит для нашей задачи вследствие отсутствия несбалансированности классов.

2 F1-score (F-мера) — среднее гармоническое precision и recall.

Где precision (точность) - доля объектов, названных классификатором положительными и при этом действительно являющимися положительными, а recall (полнота) - доля объектов положительного класса из всех объектов положительного класса нашел алгоритм.

Данная метрика, является одной из лучших в задаче бинарной классификации, так как precision не позволяет отнести все объекты к одному классу и показывает способность разделять 2 класса, recall же демонстрирует способность алгоритма обнаруживать данный класс вообще.

3 Дополнительно будем смотреть на долю верно обнаруженных здоровых и больных тканей. Так как в данной задаче обнаружение опухолевых тканей является приоритетным, хотя и не в ущерб обнаружению здоровых тканей.

7.1 Использование ML методов

В первую очередь, хочется получить некоторое наивное предсказание, которое дальше будет улучшаться. Для чего мы сначала попробуем 4 классических метода машинного обучения (Logistic Regression, SVM, Random Forest, CatBoost). Рассмотреть предсказания можно на таблице 7.1, заметим, что

Random Forest и CatBoost показали ощутимо лучшие результаты по сравнению с Logistic Regression и SVM.

	LR	SVM	RF	CB
Точность	0.893	0.898	0.927	0.927
F1 мера	0.907	0.913	0.938	0.938
Доля верно обнаруженных больных	0.863	0.887	0.919	0.919
Доля верно обнаруженных здоровых	0.938	0.914	0.938	0.938

Рис. 7.1: Предсказание классическими ML моделями

Далее на основе данных моделей создадим ансамбль (ensemble), для этого мы возьмем предсказания 4 моделей и сложим их, модели Random Forest и CatBoost пойдут с коэффициентом 2, как показавшие лучшие результаты. Таким образом, для каждого наблюдения из тестового набора данных мы имеем целочисленное предсказание в диапазоне [0, 6]. Далее мы должны в зависимости от величины числа отнести наблюдение со здоровой или больной ткани соответственно, для этого мы установим 5 порогов (0.5, 1.5, 2.5, 3.5, 4.5 - ENS_1, ENS_2, ENS_3, ENS_4, ENS_5 соответственно). На таблице 7.2 мы сравним ансамбль с методами ML, на которых мы основывались.

	LR	SVM	RF	CB	ENS_1	ENS_2	ENS_3	ENS_4	ENS_5
Точность	0.893	0.898	0.927	0.927	0.902	0.927	0.932	0.932	0.912
F1 мера	0.907	0.913	0.938	0.938	0.922	0.940	0.942	0.942	0.922
Доля верно обнаруженных больных	0.863	0.887	0.919	0.919	0.960	0.952	0.919	0.919	0.855
Доля верно обнаруженных здоровых	0.938	0.914	0.938	0.938	0.815	0.889	0.951	0.951	1.000

Рис. 7.2: Сравнение предсказаний классическими ML моделями и ансамблем, созданным на их основе

Данный ансамбль сильно улучшил изначальные результаты. Среди моделей можно выделить ENS_3 и ENS_4 показавшие лучшую F1 меру, улучшив результат CatBoost и Random Forest на 0.4%, а также ENS_5, показывающий стопроцентный шанс обнаружения больных тканей.

7.2 Использование алгоритмов кластеризации

Воспользуемся алгоритмом K-means (K-средних), будем делить наш изначальный набор данных на k кластеров, а затем, основываясь на предыдущих предсказаниях, будем относить кластеры к здоровым/больным. Были проведены исследования на количестве классов от 2 до 50, но так как на интервале [2, 15], были показаны плохие результаты, для наглядности на графике 7.3 мы покажем изменение метрик в зависимости от количества кластеров в диапазоне [15, 50].

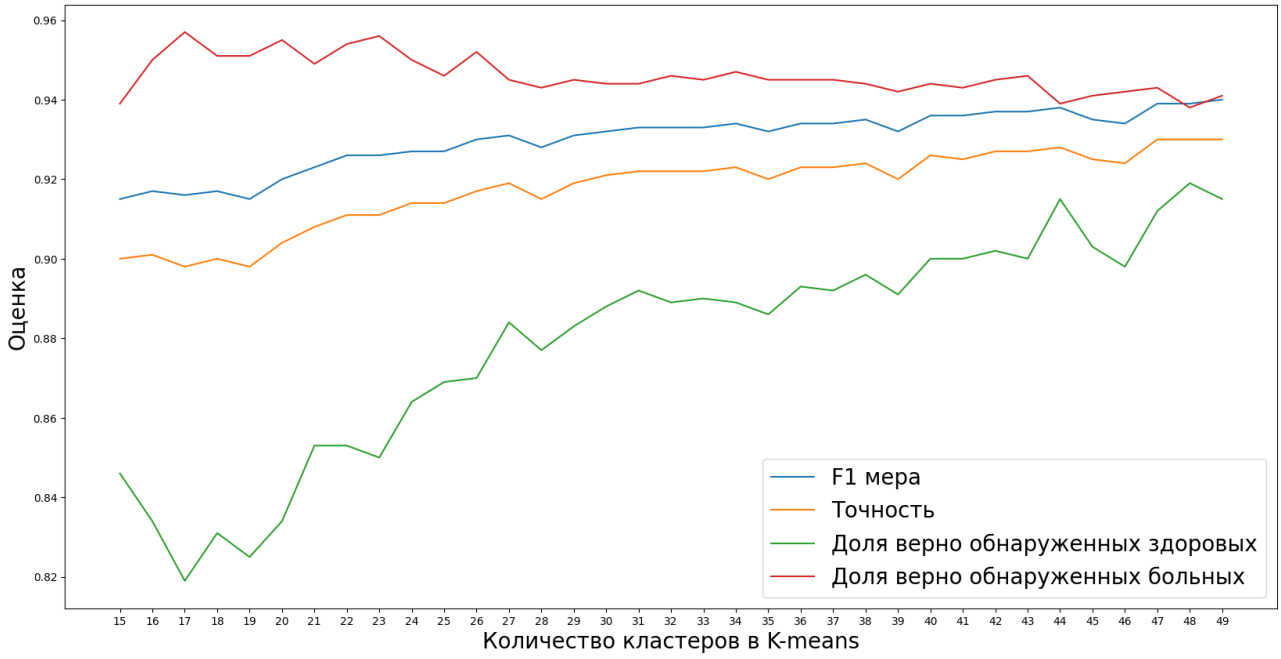

Рис. 7.3: Оценка метрик в зависимости от количества кластеров в K-means

Алгоритм не показал большого прироста результата. Только на большом количестве кластеров смогли достичь прироста на 0.2% относительно бейзлайна. Предсказания для лучших 3 разбиений представлены в таблице 7.4.

	RF	CB	ENS_3	ENS_4	K_means_1	K_means_2	K_means_3
Точность	0.927	0.927	0.932	0.932	0.930	0.930	0.930
F1 мера	0.938	0.938	0.942	0.942	0.940	0.939	0.939
Доля верно обнаруженных больных	0.919	0.919	0.919	0.919	0.941	0.938	0.943
Доля верно обнаруженных здоровых	0.938	0.938	0.951	0.951	0.915	0.919	0.912

Рис. 7.4: Сравнение предсказаний трех лучших моделей K-means с предыдущими моделями

8 Отбор признаков

Не менее важной задачей в данном исследовании будет поиск таких Рамановских сдвигов, где различие между здоровыми и больными тканями будет максимальным, с целью улучшения интерпретируемости модели и создания механизма поддержки врачебных решений. Для этого мы воспользуемся различными методами для сокращения количества признаков, с минимальной потерей качества:

1 Взаимная информация - данный метод показывает насколько чётко определена целевая переменная, если известны значения предиктора. К плюсам данного метода можно отнести то, что он позволяет находить нелинейные зависимости.

2 Exhaustive Feature Selector (Исчерпывающий набор признаков) — данный алгоритм перебирает все подмножества признаков (по мощности находящихся между установленными наименьшим и наибольшим количеством признаков), а затем выбирает тот, который показывает наилучший результат на наборе данных. Данный алгоритм очень сложно запустить при количестве признаков больше 15-20 из-за огромного времени требующегося на обработку всех подмножеств. Но он превосходит Взаимную информацию, так как анализирует признаки в совокупности, а не поодиночке.

3 Recursive Feature Elimination (Рекурсивное исключение признаков) - данный алгоритм последовательно удаляет по одному наименее полезному признаку, пока не достигнет заданного количества признаков.

4 Random Forest Importance (Метод с использованием Случайного Леса) - “обрезая” деревья ниже определенного коэффициента, мы можем подобрать наиболее важные признаки.

5 Principal Component Analysis (PCA, Метод главных компонент) - один

из главных методов, для уменьшения размерности данных, при условии потери наименьшего количества информации. Самые информативные признаки в машинном обучении — это признаки, которые показывают наибольшую изменчивость на различных объектах. Таким образом, PCA переводит большее число слабо изменчивых признаков, в меньшее число новых признаков с сильной изменчивостью, так мы понижаем размерность пространства. Основная идея PCA – уменьшить количество признаков, потеряв как можно меньше информации.

8.1 Взаимная информация

На графике 8.1 показана взаимная информация для каждого признака и целевой переменной.

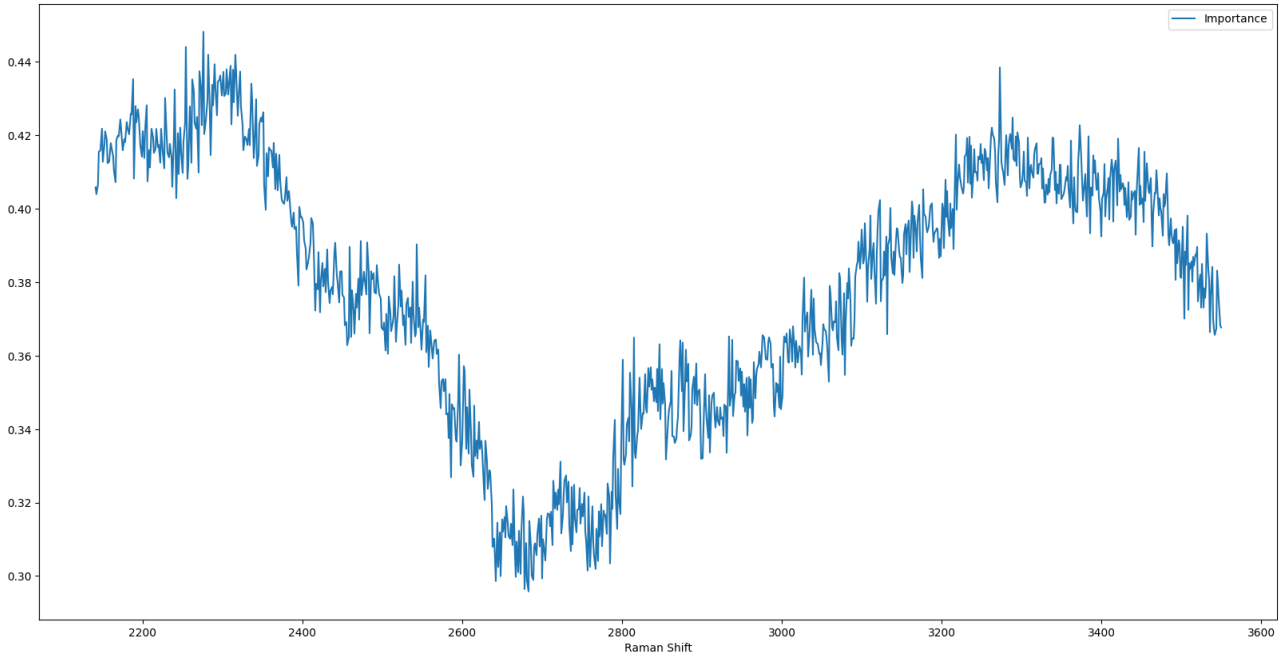

Рис. 8.1: Взаимная информация для каждого признака и целевой переменной

Заметим, что как мы и предполагали ранее, анализируя графики 6.4 и 6.5, самые важные признаки находятся в интервале [2150, 2400] cm^{-1}, также видна большая важность признаков около 3300 cm^{-1}.

Так как перед нами стоит задача максимального сокращения признаков с наименьшей потерей качества, рассмотрим результаты предсказания с помощью Random Forest на наборе данных, в зависимости от количества остав-

ленных лучших признаков. На графике 8.2 представлены результаты от 1 до 100 лучших признаков.

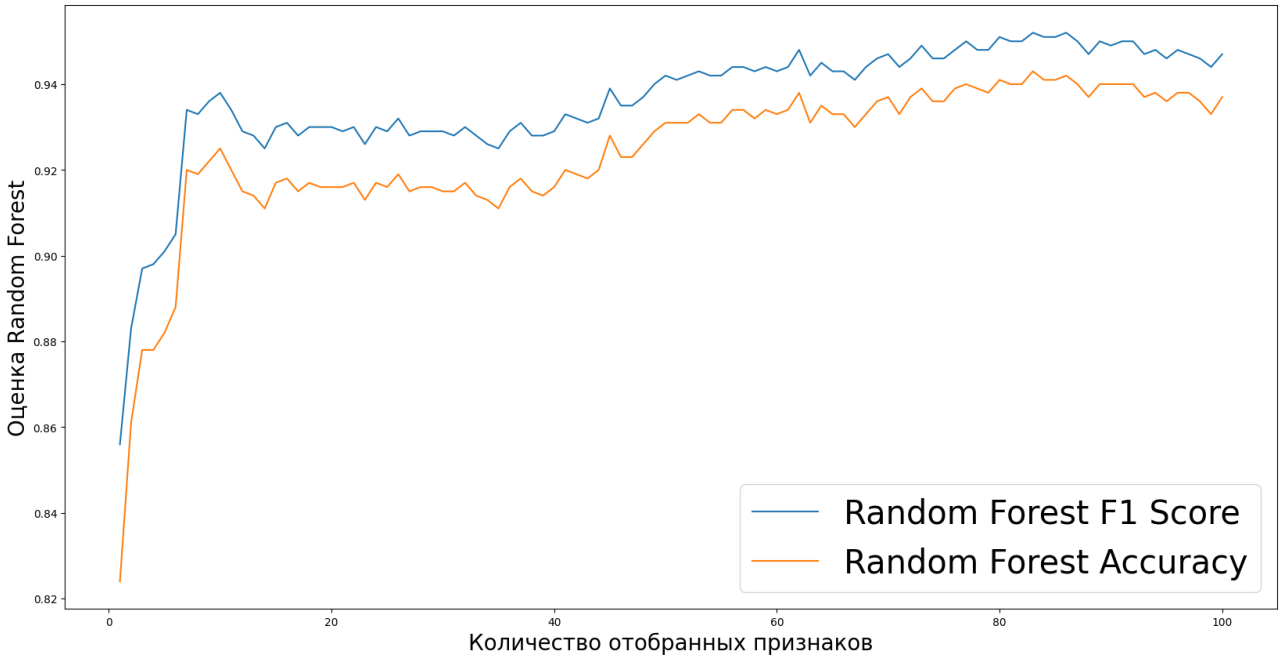

Рис. 8.2: Оценка метрик в зависимости от количества признаков от 1 до 100

Видно, что до 10 признаков метрика особо не меняется сильно. Вследствие этого на графике 8.3 покажем результаты от 5 до 15 признаков.

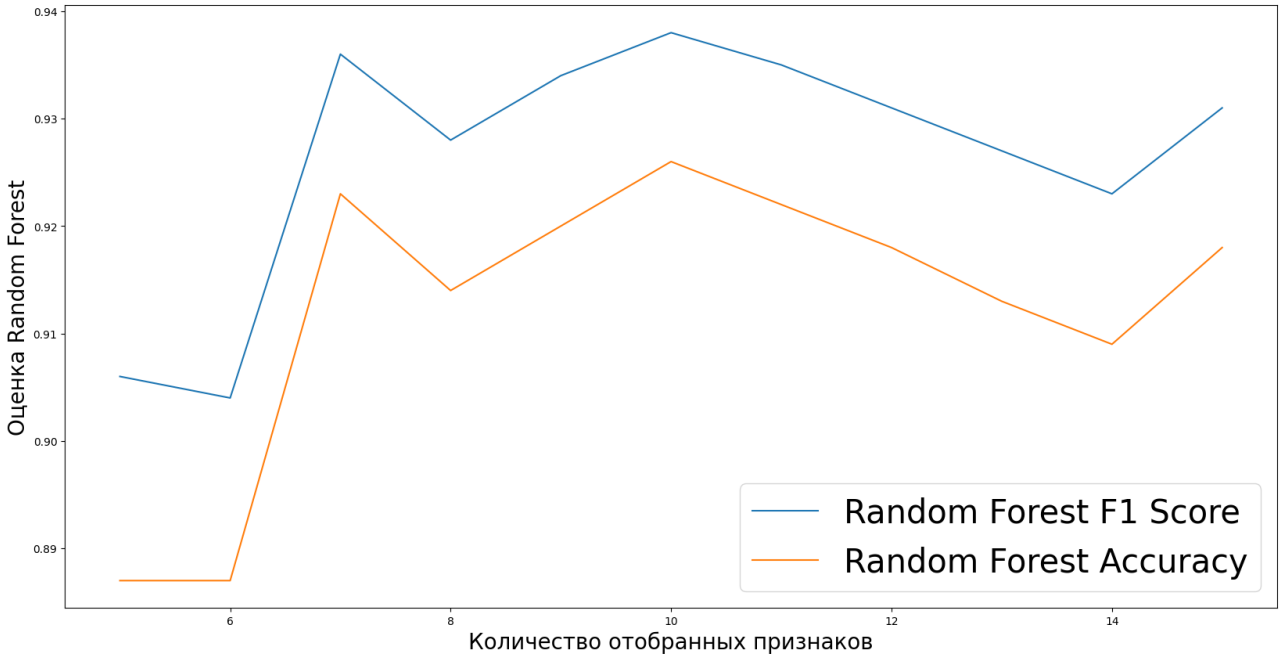

Рис. 8.3: Оценка метрик в зависимости от количества признаков от 5 до 15

Заметим, что идеальным количеством являются 7 или 10. На таблице 8.4 мы сравним результаты Random Forest, на полном наборе данных и на уре-

занном. Видно, что качество F1 меры упало лишь на 0.3% и на 1.2% относительно ансамбля моделей за счет сокращения количества признаков на 99,3% и на 99%, что является хорошим показателем.

	RF	ENS_3	ВИ_7_лучших	ВИ_10_лучших
Точность	0.927	0.932	0.917	0.927
F1 мера	0.938	0.942	0.931	0.939
Доля верно обнаруженных больных	0.919	0.919	0.927	0.927
Доля верно обнаруженных здоровых	0.938	0.951	0.901	0.926

Рис. 8.4: Сравнение предсказаний на отобранных признаках с помощью взаимной информации и бейзлайна

8.2 Exhaustive Feature Selector

Данный алгоритм превосходит взаимную информацию, за счет того, что анализирует признаки в совокупности, а не поодиночке. Запустим его для поиска лучшего подмножества от 6 до 8 признаков, среди лучших 10 признаков (EFS_1) и для поиска лучшего подмножества от 8 до 10 признаков, среди лучших 12 признаков (EFS_2). Результаты представлены в таблице 8.5, EFS_1 выбрал подмножество из 6 признаков - [2276, 2254, 2282, 2316, 3273, 2313], EFS_2 - из 9 признаков [2276, 2254, 2282, 2316, 2310, 3273, 2313, 2271, 2322].

	RF	ENS_3	ВИ_7_лучших	ВИ_10_лучших	RF_EFS_1_6_признаков	RF_EFS_2_9_признаков
Точность	0.927	0.932	0.917	0.927	0.927	0.932
F1 мера	0.938	0.942	0.931	0.939	0.939	0.943
Доля верно обнаруженных больных	0.919	0.919	0.927	0.927	0.935	0.935
Доля верно обнаруженных здоровых	0.938	0.951	0.901	0.926	0.914	0.926

Рис. 8.5: Сравнение предсказаний на основе подмножеств EFS с предыдущими моделями Random Forest

Заметим, что EFS_1 на 6 признаках показал схожие результаты с взаимной информацией на 10 признаках, а EFS_2 превзошел все предыдущие модели Random Forest, включая ансамбль моделей.

8.3 Recursive Feature Elimination

Рассмотрим результаты предсказания с помощью Random Forest на наборе данных в зависимости от количества отставленных лучших признаков. На графике 8.6 представлены результаты от 1 до 50 лучших признаков.

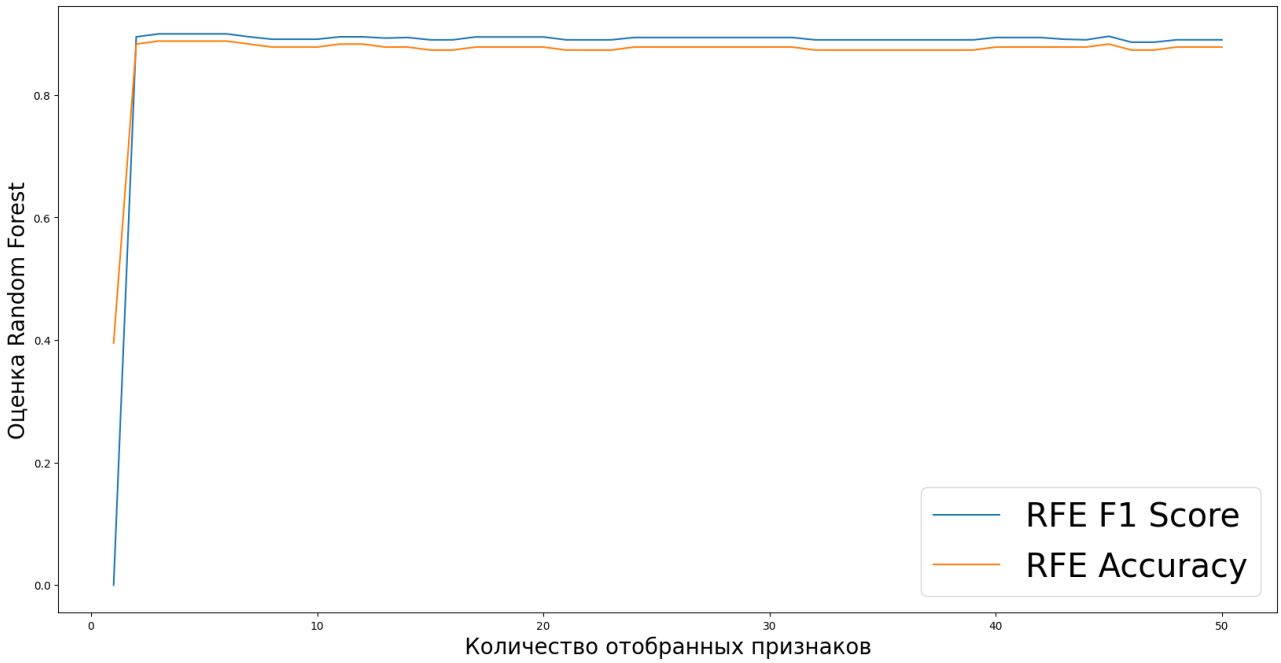

Рис. 8.6: Оценка метрик в зависимости от количества признаков от 1 до 50

Видно, что до 10 метрика особо сильно не меняется, а затем на одном отобранном признаке происходит сильное уменьшение оценок. Вследствие этого, на графике 8.8 покажем результаты от 2 до 30 признаков.

Затем отберем 10 наиболее важных признаков с помощью RFE - [2276, 2310, 2188, 2313, 2297, 2254, 2296, 2317, 2262, 3273]. На их основе построим модель. На таблице 8.7 мы сравним результаты RFE, и других методов отбора признаков.

	RF	ENS_3	RF_BИ_7	RF_BИ_10	RF_EFS_6	RF_EFS_9	RFE
Точность	0.927	0.932	0.917	0.927	0.927	0.932	0.927
F1 мера	0.938	0.942	0.931	0.939	0.939	0.943	0.940
Доля верно обнаруженных больных	0.919	0.919	0.927	0.927	0.935	0.935	0.944
Доля верно обнаруженных здоровых	0.938	0.951	0.901	0.926	0.914	0.926	0.901

Рис. 8.7: Сравнение предсказания, основанного на отборе признаков при помощи RFE, других методов отбора признаков и бейзлайна

Несмотря на то, что качество F1 меры упало на 0.3% по сравнению с алгоритмом Exhaustive Feature Selector на основе 9 признаков, оно все еще на 0.2% превосходит бейзлайн, даже при условии сокращения количества признаков на 99%.

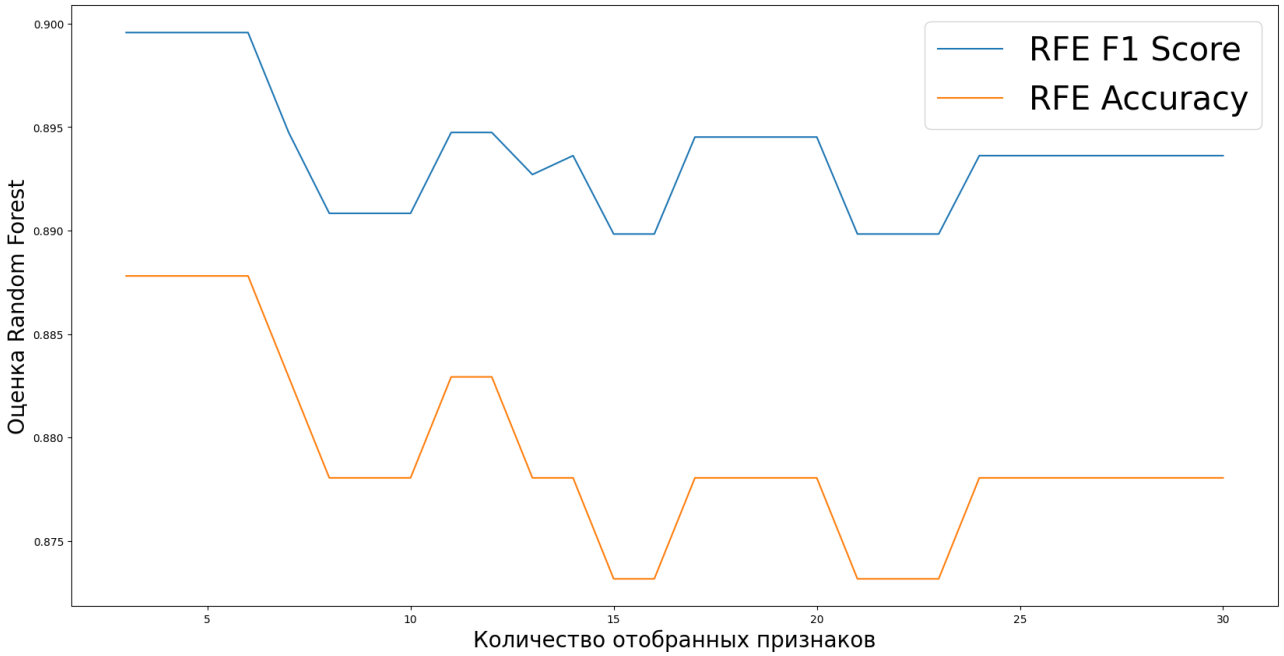

Рис. 8.8: Оценка метрик в зависимости от количества признаков от 2 до 30

8.4 Random Forest Importance

На графике 8.9 показана важность каждого из признаков, полученная с помощью "обрезания" деревьев ниже определенного коэффициента.

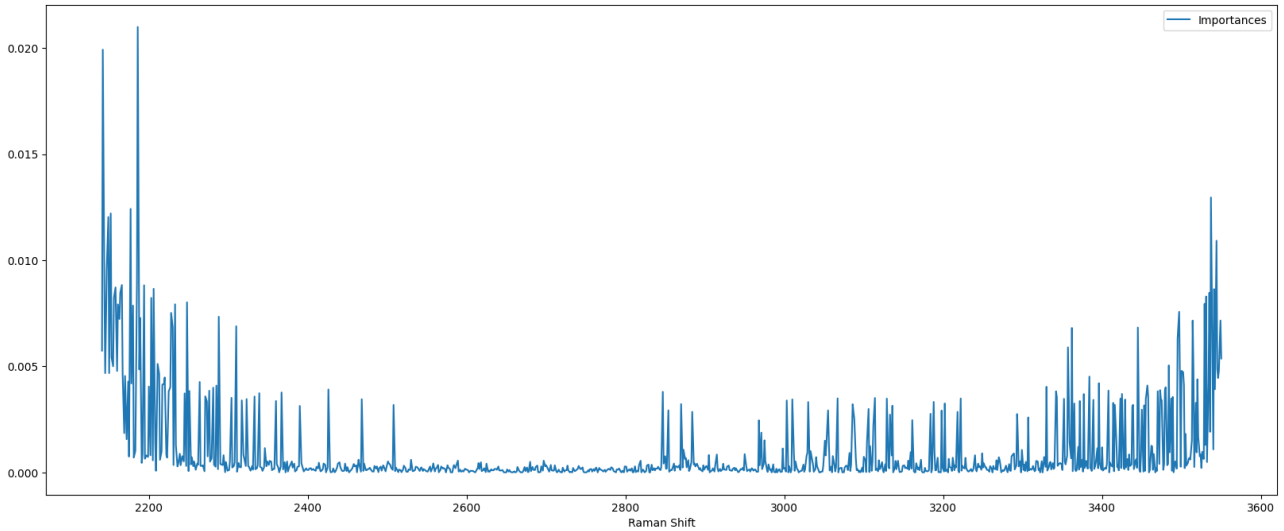

Рис. 8.9: Важность каждого признака для предсказания целевой переменной

Заметим, что признаки в интервале [2150, 2250] показывают большую важность, что совпадает с результатами полученными с помощью Взаимной информации на графике 8.1. Также видна большая важность признаков около 3500 cm^{-1}.

Рассмотрим результаты предсказания с помощью Random Forest на наборе данных, в зависимости от количества отобранных признаков. На графике 8.10 представлены результаты от 1 до 50 лучших признаков.

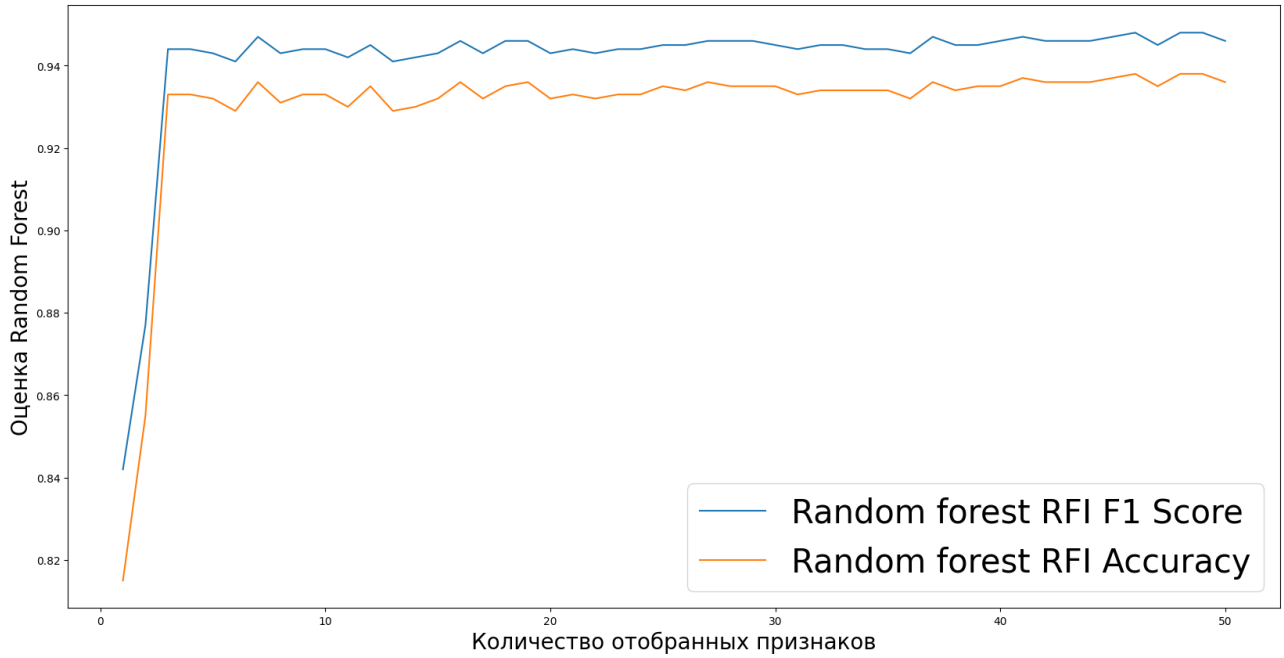

Рис. 8.10: Оценка метрик в зависимости от количества признаков от 1 до 50

Видно, что до 10 признаков метрика меняется слабо. Вследствие этого, на графике 8.11 покажем результаты от 7 до 15 признаков.

Заметим, что оптимальным количеством являются 9 признаков, в случае RFI - это [2186, 2142, 3537, 2177, 2152, 2149, 3544, 2147, 2185]. На таблице 8.12 мы сравним результаты RFI и других методов отбора признаков. Данный метод показал наилучшие результаты по сравнению с предшественниками: качество F1 меры возросло на 0.9% относительно бейзлайна при условии сокращения количества признаков на 99,1%, что является превосходным результатом.

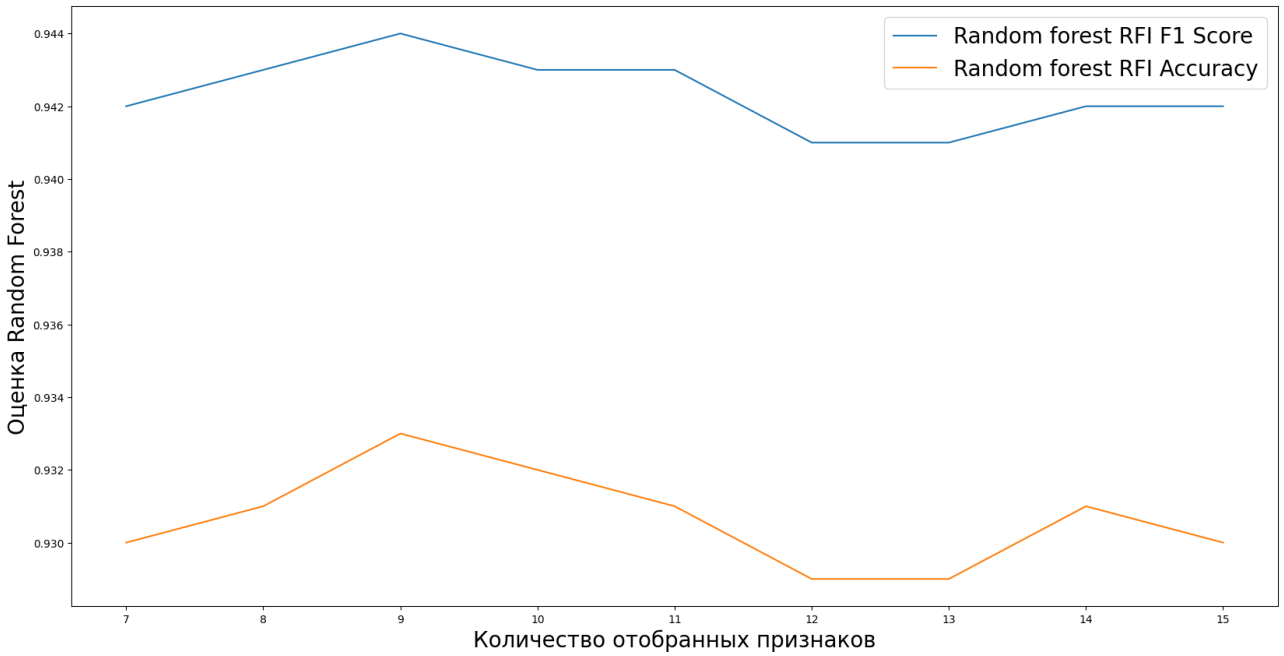

Рис. 8.11: Оценка метрик в зависимости от количества признаков от 7 до 15

	RF	ENS_3	RF_ВИ_7	RF_ВИ_10	RF_EFS_6	RF_EFS_9	RFE	RFI
Точность	0.927	0.932	0.917	0.927	0.927	0.932	0.927	0.937
F1 мера	0.938	0.942	0.931	0.939	0.939	0.943	0.940	0.947
Доля верно обнаруженных больных	0.919	0.919	0.927	0.927	0.935	0.935	0.944	0.935
Доля верно обнаруженных здоровых	0.938	0.951	0.901	0.926	0.914	0.926	0.901	0.938

Рис. 8.12: Сравнение предсказания, основанного на отборе признаков при помощи RFI, других методов отбора признаков и бейзлайна

8.5 Principal Component Analysis

Рассмотрим результаты предсказания с помощью Random Forest на датасете с обновленными признаками-компонентами в зависимости от их количества. На графике 8.13 представлены результаты преобразования от 1 до 50 компонент.

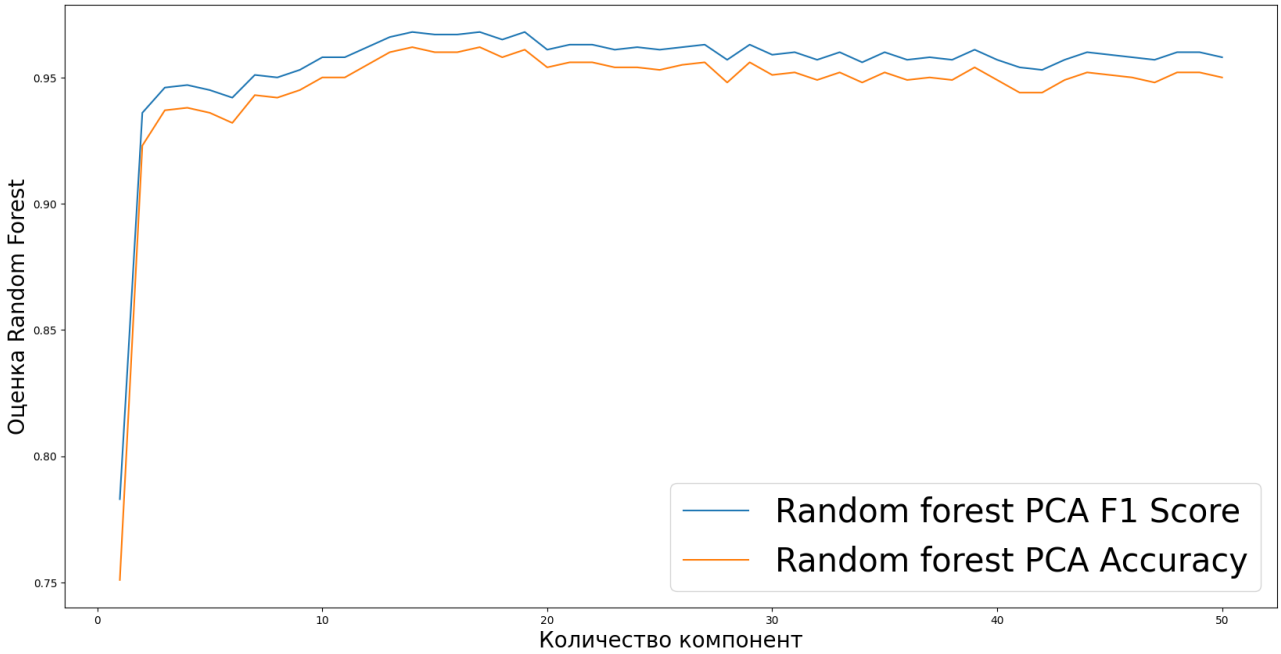

Рис. 8.13: Оценка метрик в зависимости от количества компонент в PCA

Видно, что до 20 компонент метрика почти не уменьшается, далее на 1-2 компонентах происходит резкое ухудшение результатов. Вследствие этого, на графике 8.14 покажем результаты от преобразования от 5 до 20 компонент.

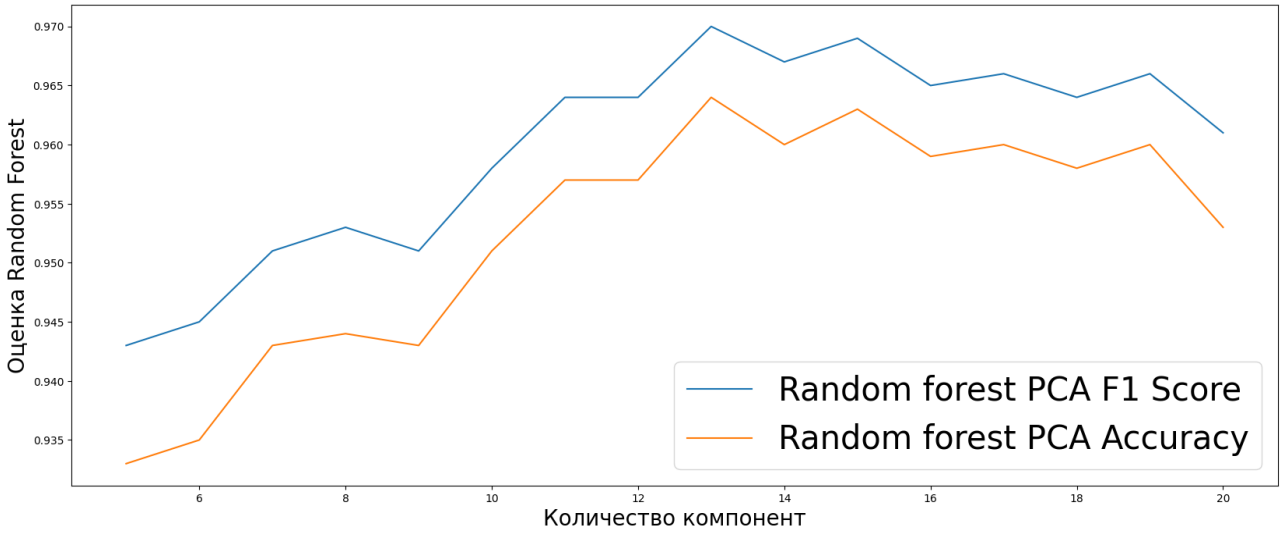

Рис. 8.14: Оценка метрик в зависимости от количества компонент в PCA

Заметим, что идеальным количеством является 8 или 13. На таблице 8.15 мы сравним результаты Random Forest, в применение к PCA, с остальными моделями на полном наборе данных. Видно, что качество F1 меры очень сильно выросло при применении метода главных компонент.

	RF	ENS_3	RF_ВИ_7	RF_ВИ_10	RF_EFS_6	RF_EFS_9	RFE	RFI	PCA_8	PCA_13
Точность	0.927	0.932	0.917	0.927	0.927	0.932	0.927	0.937	0.944	0.964
F1 мера	0.938	0.942	0.931	0.939	0.939	0.943	0.940	0.947	0.953	0.970
Доля верно обнаруженных больных	0.919	0.919	0.927	0.927	0.935	0.935	0.944	0.935	0.926	0.957
Доля верно обнаруженных здоровых	0.938	0.951	0.901	0.926	0.914	0.926	0.901	0.938	0.973	0.974

Рис. 8.15: Сравнение предсказания, основанного на уменьшении размерности признаков при помощи PCA, других методов отбора признаков и бейзлайна

8.6 Ансамбль методов

На таблице 8.15 мы видим, что самыми полезными методами оказались Exhaustive Feature Selector, Random Forest Importance, Principal Component Analysis. Объединим все признаки от данных моделей, получим 13 признаков, полученных с помощью PCA, а также 18 признаков - [2142, 2147, 2149, 2152, 2177, 2185, 2186, 2254, 2271, 2276, 2282, 2310, 2313, 2316, 2322, 3273, 3537, 3544], полученные при объединении признаков EFS и RFI. На таблице 8.16 мы сравним результаты Random Forest в применение к ансамблю, с остальными моделями на полном наборе данных. Видно, что качество F1 меры при объединении EFS, RFI и PCA сильно выросло и превзошло все методы по отдельности, а также дало 2.5 кратный прирост к результату бейзлайна.

	RF	ENS_3	RF_ВИ_7	RF_ВИ_10	RF_EFS_6	RF_EFS_9	RFE	RFI	PCA_8	PCA_13	EFS_RFI_PCA
Точность	0.927	0.932	0.917	0.927	0.927	0.932	0.927	0.937	0.944	0.964	0.971
F1 мера	0.938	0.942	0.931	0.939	0.939	0.943	0.940	0.947	0.953	0.970	0.975
Доля верно обнаруженных больных	0.919	0.919	0.927	0.927	0.935	0.935	0.944	0.935	0.926	0.957	0.967
Доля верно обнаруженных здоровых	0.938	0.951	0.901	0.926	0.914	0.926	0.901	0.938	0.973	0.974	0.976

Рис. 8.16: Сравнение предсказаний классическими ML моделями и ансамблем, созданным на их основе

9 Заключение

В результате работы были получены выводы об эффективности использования различных методов бинарной классификации тканей. Случайный лес и Catboost показали лучшие результаты среди моделей классического машинного обучения. На основе случайного леса и методов отбора признаков (Exhaustive Feature Selector, Random Forest Importance и Principal Component Analysis) в дальнейшем были достигнуты лучшие результаты. В итоге была получена модель, классифицирующая ткани с помощью Рамановской спектроскопии на здоровые и больные в 2.5 раза лучше бейзлайна (рост качества F1 меры с 0.938 до 0.975).

Также были произведен отбор наиболее важных Рамановских сдвигов, с целью улучшения интерпретируемости модели и создания механизма поддержки врачебных решений. Качество F1 меры при применении алгоритмов Exhaustive Feature Selector и Random Forest Importance возросло на 0.5% и на 0.9% соответственно при условии сокращения количества признаков на 99,1% (с 1015 до 9), что является отличным показателем. Также при применении Principal Component Analysis была получена F1 мера = 0.97, что дало двукратный прирост к метрике бейзлайна.

Тем не менее надо понимать, что диагностика опухолевых тканей - сложная задача, и ее контекст зависит как от страны, где расположена лаборатория, так и от месторасположения тканей. Так, полученные в результате работы выводы, могут считать бейзлайном для дальнейших предсказаний в будущем, но не отменяют необходимость исследований с учетом нового контекста.

Список литературы

Monika Gniadecka, Peter Alshede Philipsen, Sigurdur Sigurdsson, Sonja Wessel, Ole Faurskov Nielsen, Daniel Højgaard Christensen, Jana Hercogova, Kristian Rossen, Henrik Klem Thomsen, Robert Gniadecki, Lars Kai Hansen, and Hans Christian Wulf. Melanoma diagnosis by raman spectroscopy and neural networks: structure alterations in proteins and lipids in intact cancer tissue. *Journal of Investigative Dermatology, vol. 122, no. 2, pp. 443–449*, 2004. URL https://pubmed.ncbi.nlm.nih.gov/15009728/.

Zhiwei Huang, Annette McWilliams, Harvey Lui, David I McLean, Stephen Lam, and Haishan Zeng. Near-infrared raman spectroscopy for optical diagnosis of lung cancer. *International Journal of Cancer, vol. 107, no. 6, pp. 1047–1052*, 2003. URL https://pubmed.ncbi.nlm.nih.gov/14601068/.

Zhiwei Huang, Harvey Lui, David I McLean, Mladen Korbelik, and Haishan Zeng. Raman spectroscopy in combination with background near-infrared autofluorescence enhances the in vivo assessment of malignant tissues. *Photochemistry and Photobiology, vol. 81, no. 5, p. 1219*, 2005. URL https://pubmed.ncbi.nlm.nih.gov/33802369/.

Michael Jermyn, Joannie Desroches, Jeanne Mercier, Marie-Andrée Tremblay, Karl St-Arnaud, Marie-Christine Guiot, Kevin Petrecca, and Frederic Leblond. Neural networks improve brain cancer detection with raman spectroscopy in the presence of operating room light artifacts. *Journal of Biomedical Optics, vol. 21, no. 9, 094002*, 2016. URL https://pubmed.ncbi.nlm.nih.gov/27604560/.

Wooje Lee, Aufried T.M. Lenferink, Cees Otto, and Herman L. Offerhaus. Classifying raman spectra of extracellular vesicles based on convolutional neural networks for prostate cancer detection. *Journal of Raman Spectroscopy, vol. 51, no. 2, pp. 293–300*, 2019. URL https://www.researchgate.net/publication/337106643_Classifying_

Raman_spectra_of_extracellular_vesicles_based_on_convolutional_neural_networks_for_prostate_cancer_detection.

Marco Riva, Tommaso Sciortino, Riccardo Secoli, Ester D'Amico, Sara Moccia, Bethania Fernandes, Marco Conti Nibali, Lorenzo Gay, Marco Rossi, Elena De Momi, and Lorenzo Bello. Glioma biopsies classification using raman spectroscopy and machine learning models on fresh tissue samples. *Cancers, vol. 13, no. 5, p. 1073*, 2021. URL https://pubmed.ncbi.nlm.nih.gov/33802369/.

Stephan Seifert. Application of random forest based approaches to surface-enhanced raman scattering data. *Scientific Reports, vol. 10, no. 1*, 2020. URL https://pubmed.ncbi.nlm.nih.gov/32214194/.

Linwei Shang, Jinlan Tang, Jinjin Wu, Hui Shang, Xing Huang, Yilin Bao, Zhibing Xu, Huijie Wang, and Jianhua Yin. Polarized micro-raman spectroscopy and 2d convolutional neural network applied to structural analysis and discrimination of breast cancer. *Biosensors, vol. 13, no. 1, p. 65*, 2022. URL https://pubmed.ncbi.nlm.nih.gov/36671896/.

Effendi Widjaja, Wei Zheng, and Zhiwei Huangg. Classification of colonic tissues using near-infrared raman spectroscopy and support vector machines. *International Journal of Oncology, vol. 32, no. 3, pp. 653–662*, 2008. URL https://pubmed.ncbi.nlm.nih.gov/18292943/.