Санкт-Петербургский государственный университет

Белавин Сергей Андреевич

Выпускная квалификационная работа

Использование методов машинного обучения для анализа данных в сложной медико-биологической системе

Уровень образования: магистратура
Направление 01.04.02 «Прикладная математика и информатика»
Основная образовательная программа ВМ.5691.2019 «Прикладная математика и информатика в задачах медицинской диагностики»

Научный руководитель:
д-р техн. наук,
профессор кафедры КММС,
Дегтярев Александр Борисович

Рецензент:
к.м.н., ведущий научный сотрудник
ФГБУ НМИЦ ПН им. В.М. Бехтерева» Минздрава РФ,
Залуцкая Наталья Михайловна

Санкт-Петербург
2021

Оглавление

Введение

Предметная область: Сложность задач, возникающих в области нейрофизиологии, а также объем данных, которые используются для диагностики и лечения пациентов, неуклонно растут. В связи с этим разработка эффективных методов обработки этих данных становится наиболее приоритетной. В настоящее время медицинские исследования, и, в частности, нейрофизиология, являются обширной областью для применения методов искусственного интеллекта.

По оценкам, в 2018 году во всем мире насчитывалось около 50 миллионов человек, страдающих деменцией. К 2030 году это число по прогнозам увеличится до 82 миллионов, а уже к 2050 достигнет отметки 152 миллиона человека, что почти в 3 раза больше, чем на данный момент [1].

Болезнь Альцгеймера — это тип деменции, составляющий около 60-80% всех случаев, который вызывает проблемы с памятью, мышлением и поведением. Симптомы обычно развиваются медленно и со временем ухудшаются, становясь достаточно серьезными, чтобы мешать повседневным задачам. Повреждение мозга начинается задолго до того, как появляются проблемы с памятью или другие когнитивные проблемы [2].

Для диагностики болезни Альцгеймера требуется комплексное обследование: изучается семейная медицинская история, применяются различные методы нейровизуализации, например магнитно-резонансная томография (МРТ), проводятся нейрофизиологические исследования, например электроэнцефалография (ЭЭГ), когнитивные тесты на оценку памяти и мышления, анализы крови [3].

Актуальность исследования: Ранняя диагностика болезни Альцгеймера может позволить назначить своевременное лечение для замедления прогрессирования заболевания. Так как, на практике приходится вручную анализировать полученные результаты обследований, возможны

ошибки, связанные с человеческим фактором. Кроме того, не всегда удается поставить точный диагноз. Учитывая большое количество источников данных, необходимо знать, на какие показатели стоит обращать внимание.

В данной работе будут рассмотрены данные полученные с помощью МРТ и ЭЭГ обследований, теста Векслера, а также анализов крови.

Постановка задачи

Целью данной работы является исследование возможности разделения пациентов на группы, используя данные результатов теста Векслера, а также данные МРТ и ЭЭГ обследований головного мозга.

Для выполнения поставленной задачи были выделены следующие этапы:

1. **Формирование набора данных.** Извлечение необходимых для проведения исследования данных из обследований пациентов. Совмещение данных из разных источников.
2. **Предобработка данных.** Удаление пропущенных значений и нормирование.
3. **Проведение анализа данных.** Отбор наиболее информативных признаков. Проведение кластерного анализа. Поиск различий между выделенными группами по показателям анализов крови. Оценка полученных результатов.

Обзор литературы

В большинстве работ, для диагностирования болезни Альцгеймера и других заболеваний мозга, используют результаты только одного вида обследования пациентов. Наиболее показательными и информативными являются МРТ головного мозга [4-8], ЭЭГ [9-10], тестирование психического/интеллектуального состояния пациента [11-13], а также анализы крови [14, 15].

Реже в исследованиях встречается использование сразу нескольких источников данных. Так в [11] используются данные о когнитивном состоянии, анализы спинномозговой жидкости и морфометрических параметров МРТ изображений.

Анализ снимков МРТ осложняется тем, что от аппарата, на котором эти снимки были сделаны, а также от их качества, могут зависеть результаты обработки и сегментации этих снимков. Однако, для упрощения анализа снимков, можно использовать параметры, полученные с помощью программных средств сегментации, таких как FreeSurfer [4-8].

Классическим методом предобработки электроэнцефалограмм является извлечение некоторых волновых характеристик зарегистрированных сигналов. Так, электроэнцефалограмма рассматривается как многоканальный сигнал, каждый из каналов которого имеет собственные спектральные характеристики. В работе [16] проведен анализ достоинств и недостатков каждой из них. В рамках применения машинного обучения для диагностирования заболеваний мозга, в [9] используется когерентность сигналов, а в [10] спектральная мощность.

Зачастую, размерность признакового пространства является большой, и для анализа данных, а также визуализации, используются различные методы понижения размерности. Одним из основных методов ее понижения можно считать метод главных компонент (PCA) [7]. Так же в работе [7] для

понижения размерности рассматривался метод линейного дискриминантного анализа (LDA), который позволил получить более точные результаты, однако требует наличия известных меток классов. В работе [17] был проведен анализ применимости методов PCA, t-SNE (Стохастическое вложение соседей с t-распределением) и LDA для понижения размерности многомерных медицинских данных. Был сделан вывод, что снижение размерности признакового пространство позволяет получить более точные результаты кластеризации.

В исследовании [15] проводился анализ параметров окислительного стресса при психических нарушениях. Было выявлено различие у пациентов с болезнью Альцгеймера по показателю супероксиддисмутазы (СОД). В исследовании [13] проводилось сравнение показателей теста Векслера при болезни Альцгеймера и легких когнитивных нарушениях со здоровыми пациентами, и был сделан вывод о том, что наиболее сильно страдают регионы мозга, отвечающие за вербальные навыки.

Подавляющее большинство исследований сосредотачивается на попытке построить классифицирующий алгоритм, однако работы, в которых ставится задача кластеризации пациентов, у которых отсутствуют точные диагнозы, практически нет. В частности, нет подобных работ, занимающихся анализом сразу нескольких видов обследований пациентов.

Глава 1. Исходные данные

1.1 Описание данных

Для проведения данного исследования Санкт-Петербургским научно-исследовательским психоневрологическим институтом имени В. М. Бехтерева были предоставлены анонимизированные данные обследований пациентов, страдающих некоторыми заболеваниями головного мозга, в том числе болезнью Альцгеймера, а также контрольную группу.

Количество пациентов для различных обследований:

- Тест Векслера и анализы крови: 199 человек;
- ЭЭГ: 151 человека;
- МРТ: 117 человек.

Так же для каждого пациента был предоставлен предположительный, неточный диагноз (метка):

1. Болезнь Альцгеймера;
2. Сосудистая деменция;
3. Депрессия;
4. Контрольная группа.

1.2 Тест Векслера

Тест Векслера является одним из самых известных тестов для измерения уровня интеллектуального развития. Он был разработана Дэвидом Векслером в 1939 году. Тест основан на иерархической модели интеллекта Д. Векслера и диагностирует общий интеллект и его составляющие - вербальный и невербальный интеллекты [18].

Для каждого пациента предоставлен набор показателей (рис. 1), соответствующих различным субтестам, и итоговому показателю теста:

- Субтест I «Личные и общественные данные»;
- Субтест II «Ориентировка»;
- Субтест III «Психический контроль»;

- Субтест IV «Логическая память»;
- Субтест V «Цифры»;
  - Субтест V(a) «Цифры – прямой порядок»;
  - Субтест V(b) «Цифры – обратный порядок»;
- Субтест VI «Зрительная ретенция»;
- Субтест VII «Парные ассоциации»;
  - Субтест VII(a) «Субтест – лёгкие пары»;
  - Субтест VII(b) «Субтест – сложные пары»;
- ЭИПП (Эквивалентный интеллекту показатель памяти).

id	B I	B II	B III	B IV	B V	B Va пр	B Vb обр	B VI	B VII	B VIIa легк	B VIIb трудн	B ЭИПП
1	5	5	3	8	11	6	5	4	13	9	4	112
2	6	5	4	7	10	6	4	2	10	7	3	94
3	6	5	6	11	11	6	5	11	21	9	12	143
4	6	5	5	11	12	6	6	6	18	9	9	120
5	6	4	4	4	9	6	3	1	5	5	0	81
6	6	5	4	6	11	6	5	5	17	9	8	122

Рис. 1. - Пример данных теста Векслера

1.3 Данные МРТ

Результаты МРТ обследований пациентов были предоставлены в виде текстовых файлов, полученных при обработке снимков программным пакетом FreeSurfer [19]. Данный программный пакет для различных структур мозга вычисляет необходимые для исследования параметры, в нашем случае для указанных выше заболеваний врачи учитывают следующие значения:

- количество белого вещества;
- количество серого вещества;
- средняя толщина коры;
- площадь поверхности.

Эти параметры вычисляются для различных долей (лобной, теменной, затылочной, височной, инсулы и гиппокампа) левого и правого полушарий

мозга.

Для извлечения информации из текстовых файлов был написан парсер, с помощью которого все результаты были сведены в общую таблицу (рис. 2).

wm-lh-isthmuscingulat	wm-lh-lateraloccipita	wm-lh-lateralorbitofronta	wm-lh-lingual	wm-lh-medialorbitofrontal
2868	9104	5780	5847	2910
3698,5	7181,7	6396,6	4962,2	3472,6
4888,9	10487,8	6465	5958,2	3745,9
3799,3	8575	6340,2	6291,1	3835,5
3722,6	7842,7	5255,5	4429,7	3209,9
3604,1	6650,2	5525,8	4473,4	3977,8
5009,5	9662,7	7766,9	6372,5	3579,5
3177,3	6645,2	5316,4	3508,1	2468,3
3801,7	10187,3	7096,3	6372,1	4042
4461,3	10134,8	7827,4	6599,6	3798,6
3394,3	9695,3	5888,8	5754,3	3321,4

Рис. 2. - Пример части таблицы данных МРТ

1.4 Данные ЭЭГ

Электроэнцефалограммы пациентов были записаны на аппарате МИЦАР ЭЭГ и представлены в виде файлов в формате .eeg (рис. 3). Запись происходила в состоянии покоя и при фотостимуляции.

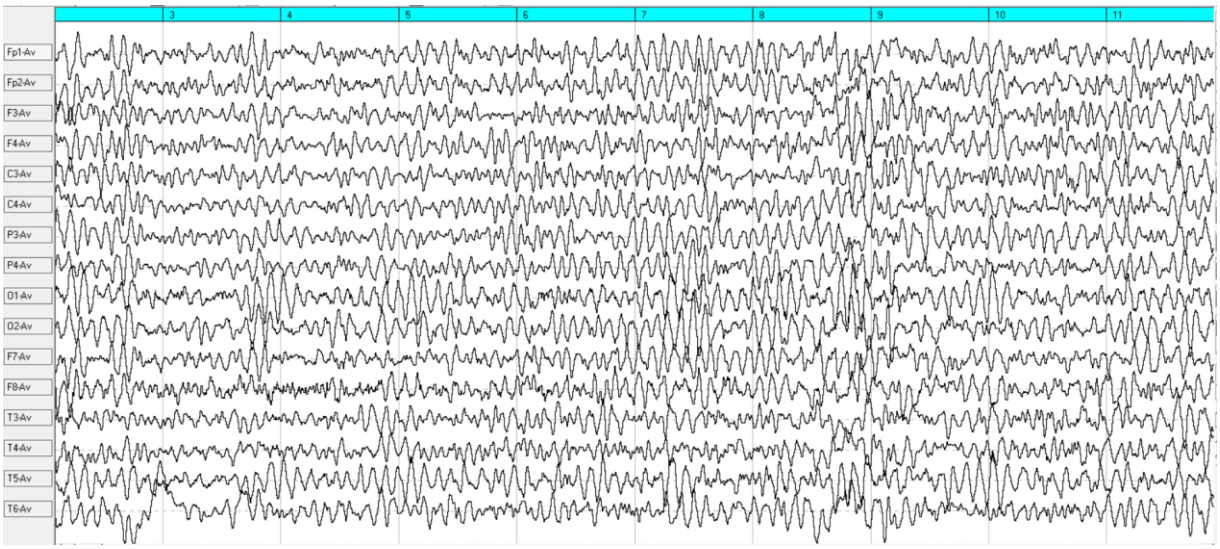

Рис. 3. - Фрагмент записи ЭЭГ

Существуют различные характеристики, которые можно извлечь из записи ЭЭГ для анализа. Однако, экспертами из НИПНИ им. Бехтерева было рекомендовано использовать когерентный анализ, который считается индикатором функциональных взаимосвязей между различными корковыми областями [20, 21].

Для приведения данных в общий вид, подходящий для исследований, а также для проведения когерентного анализа был использован программный пакет WinEEG.

Процесс извлечения необходимых признаков состоял из следующих этапов:

- приведение записи к усредненному монтажу (с параметрами: фильтр низких частот – 30 Гц, фильтр высоких частот – 0.3 Гц, режекторный фильтр – 50 Гц) по 16 каналам (Fp1, Fp2, F7, F3, F4, F8, T3, C3, C4, T4, T5, P3, P4, T6, O1, O2);
- выбор участка записи без артефактов в состоянии покоя (в режиме фоновой ЭЭГ) длиной в 15 секунд;
- исключение шумов встроенными инструментами WinEEG;
- извлечение необходимых признаков с помощью когерентного анализа.

Для каждого канала были получены четыре значения когерентности, соответствующие дельта-, тета-, альфа- и бета-ритмам головного мозга. Полученные данные были сведены в общую таблицу (рис. 4).

Fp1-Av Del	Fp2-Av Del	F3-Av Del	F4-Av Del	C3-Av Del	C4-Av Del	P3-Av Del	P4-Av Del	O1-Av Del
0,253	0,26	0,333	0,219	0,335	0,282	0,3	0,324	0,267
0,28	0,318	0,274	0,295	0,265	0,249	0,325	0,275	0,324
0,307	0,302	0,26	0,24	0,221	0,168	0,257	0,216	0,328
0,306	0,343	0,334	0,349	0,339	0,248	0,399	0,275	0,393
0,148	0,217	0,211	0,205	0,196	0,23	0,253	0,213	0,24
0,27	0,319	0,246	0,287	0,215	0,192	0,297	0,31	0,337
0,294	0,299	0,263	0,221	0,225	0,239	0,3	0,294	0,317
0,367	0,354	0,288	0,317	0,206	0,27	0,355	0,419	0,397
0,246	0,324	0,282	0,324	0,23	0,265	0,298	0,273	0,3
0,171	0,283	0,245	0,312	0,282	0,26	0,282	0,226	0,308
0,222	0,236	0,258	0,229	0,279	0,259	0,322	0,269	0,283
0,277	0,298	0,312	0,332	0,254	0,254	0,214	0,297	0,322
0,138	0,448	0,35	0,453	0,329	0,455	0,404	0,357	0,388
0,254	0,291	0,269	0,288	0,324	0,317	0,311	0,341	0,253
0,258	0,275	0,257	0,274	0,179	0,192	0,28	0,268	0,308
0,301	0,346	0,278	0,297	0,332	0,286	0,376	0,357	0,368
0,276	0,338	0,332	0,35	0,316	0,302	0,355	0,36	0,366
0,366	0,427	0,365	0,456	0,364	0,461	0,47	0,433	0,427
0,314	0,231	0,453	0,448	0,472	0,459	0,435	0,454	0,376
0,247	0,295	0,299	0,313	0,229	0,249	0,338	0,312	0,331
0,356	0,306	0,375	0,287	0,34	0,303	0,298	0,398	0,306
0,416	0,376	0,345	0,371	0,323	0,371	0,44	0,427	0,446
0,488	0,453	0,492	0,425	0,414	0,503	0,405	0,526	0,469
0,321	0,312	0,272	0,307	0,241	0,241	0,364	0,37	0,368
0,238	0,271	0,206	0,262	0,199	0,223	0,254	0,295	0,281

Рис. 4. - Пример части таблицы данных ЭЭГ

1.5 Анализы крови

Список показателей крови, рассматриваемых в исследовании, представлен в таблице 1.

Цитогенетика	Биохимия	Общий анализ	Гормоны	Маркеры оксидативного стресса
Цитогенетика	Альбумин Холестерин Общий белок Глюкоза ЛПВП ЛПНП Триглицериды С-реактивный белок Коэффициент атерогенности	Hb Эритроциты Тромбоциты Лейкоциты СОЭ Ретикулоциты	Пролактин Кортизол	СОД Е/мл СОД Е/г Hb ГР в эритроцитах Е/г Hb ГР в плазме Е/л ГП в эритроцитах Е/л Каталаза

Таблица 1. - Список показателей крови

1.6 Предобработка данных

Все полученные данные были объединены в общую таблицу по уникальному идентификационному номеру пациента.

Для получения более качественных результатов данные МРТ обследований были нормированы отдельно для каждого пола, т.к. размер долей мозга у разных полов может отличаться.

Значения столбцов всех пациентов были преобразованы по следующей формуле:

$$X = \frac{X - X_{min}}{X_{max} - X_{min}},$$

где X_{max} – максимальное значение в столбце, а X_{min} – минимальное.

Глава 2. Используемые методы

2.1 Кластеризация

Общая формулировка задачи кластеризации звучит следующим образом. Дано множество объектов X и множество номеров кластеров Y. Задана некоторая функция расстояния между объектами $\rho = \rho(x, x')$. Необходимо разбить объекты по кластерам таким образом, чтобы расстояние между объектами одного кластера было небольшим, а между объектами из разных кластеров существенно отличалось [22].

Дальше будут рассмотрены методы кластеризации, которые были применены в данном исследовании.

2.1.1 K-средних

Простейшим методом кластеризации является метод k-средних (k–means) [23], где k — число кластеров, которые требуется выделить. Предполагается, что число k известно.

В начале работы алгоритма выбираются k случайных точек μ_i. Это точки играют роль центров кластеров. После этого выполняются следующие итерации (рис. 5).

Каждая точка x_i будет отнесена к тому кластеру Y_j, расстояние до центра которого минимально:

$$j = \arg\min_{\mathrm{k}} \rho(x_i, \mu_k)^2$$

На каждом шаге в качестве нового центра кластера выбирается среднее арифметическое всех точек, попавших в этот кластер, и обновляются метки кластеров для точек в зависимости от близости к новым центрам кластеров.

$$\mu_i = \frac{1}{|Y_j|} \sum_{x \in Y_j} x$$

Алгоритм завершается, когда на итерации не происходит изменения внутрикластерного расстояния.

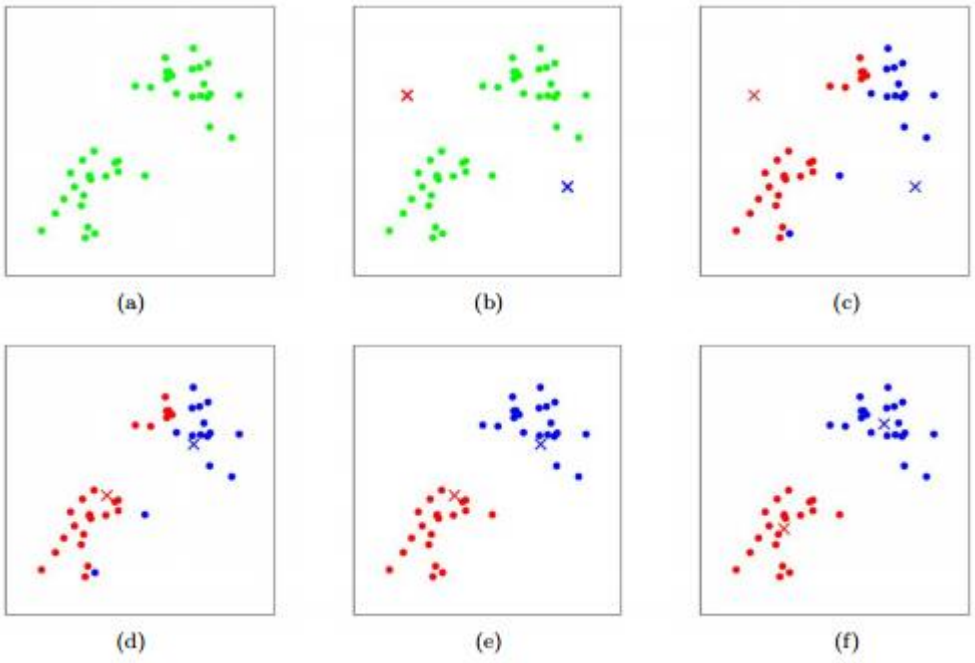

Рис. 5. - Пример работы алгоритма k-средних [24].

2.1.2 Агломеративная кластеризация

Агломеративная кластеризация – алгоритм кластеризации, направленный на построение вложенных разбиений исходного множества [23]. Новые кластеры создаются объединением более мелких.

Алгоритм агломеративной кластеризации:

- Изначально каждый объект содержится в собственном кластере.
- Итеративно: два ближайших (на основании выбранной функции расстояния) кластера сливаются.
- Процесс продолжается, пока не останется необходимое кол-во кластеров.

Функции расстояния между кластерами Y и Z [22]:

- Расстояние ближнего соседа: $\rho(Y,Z) = \min_{y\in Y, z\in Z} \rho(y,z)$
- Расстояние дальнего соседа: $\rho(Y,Z) = \max_{y\in Y, z\in Z} \rho(y,z)$
- Среднее расстояние: $\rho(Y,Z) = \frac{1}{|Z||Y|}\sum_{y\in Y}\sum_{z\in Z}\rho(y,z)$
- Центроидное расстояние: $\rho(Y,Z) = \rho^2(\sum_{y\in Y}\frac{y}{|Y|}, \sum_{z\in Z}\frac{z}{|Z|})$

– Расстояние Уорда: $\rho(Y,Z) = \frac{|Z||Y|}{|Z|+|Y|}\rho^2(\sum_{y\in Y}\frac{y}{|Y|}, \sum_{z\in Z}\frac{z}{|Z|})$

Данный метод позволяет строить деревья объединения кластеров (дендрограммы (рис. 6)), что позволяет удобно определять оптимальное количество кластеров.

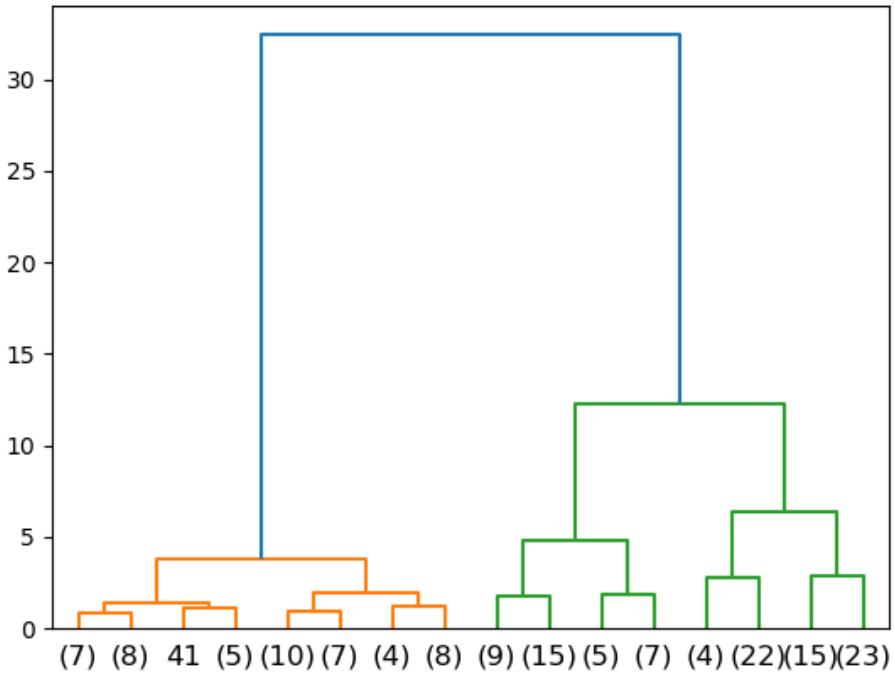

Рис. 6. - Пример построения дендрограммы [25]

2.1.3Модель гауссовой смеси

Модель гауссовой смеси [26] представляет собой EM-алгоритм [22] для гауссовых распределений.

EM – алгоритм заключается в максимизации правдоподобия. Он основан на том, что плотность вероятности распределения точек $p(x)$ выборки представляет собой взвешенную сумму плотностей вероятности $p_j(x)$ в каждом кластере

$$p(x) = \sum_{j=1}^{K} w_j p_j(x)$$

В нашем случае $p_j(x)$ выбираются из гауссового распределения.

EM-алгоритм реализуется выполнением двух шагов:

E-шаг: вычисляются вспомогательные переменные:

$$g_{ji} = p(\theta_j|x_i) = \frac{w_j p_j(x_i)}{p(x_i)}$$

Эти параметры фиксируются.

М-шаг: при зафиксированных g_{ji} решение задачи максимизации правдоподобия может быть найдено согласно:

$$w_j = \frac{1}{N}\sum_{i=1}^{N} g_{ji},$$

$$\theta_j = \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{N} g_{ji} ln\varphi(\theta; x)$$

Объект относится к кластеру j, для которого максимально значение $p(\theta_j|x_i)$.

Итерации происходят до сходимости.

2.2 Понижение размерности

В связи с тем, что признаковое пространство велико по сравнению с количеством объектов, а также для удобства визуализации данных, необходимо уменьшить количество признаков, при этом не потеряв информацию. Одним из методов, который позволяет сделать это, является метод главных компонент.

В анализе главных компонент строится новое признаковое пространство, которое с помощью линейного преобразования способно восстановить исходное с минимальными потерями информации [22].

Предположим, имеется k объектов, каждый из которых описан n признаками:

$$x_i = \big(f_1(x_i), \dots, f_n(x_i)\big), \qquad i = \overline{1, \dots, k}$$

В матричном виде:

$$F = \begin{pmatrix} f_1(x_1) & \cdots & f_n(x_1) \\ \vdots & \ddots & \vdots \\ f_1(x_k) & \cdots & f_n(x_k) \end{pmatrix}$$

Новое признаковое пространство размерности $m < n$ тех же объектов обозначим:

$$z_i = \big(g_1(x_i), \dots, g_m(x_i)\big), \qquad i = \overline{1, \dots, k}$$

В матричном виде:

$$G = \begin{pmatrix} g_1(x_1) & \cdots & g_m(x_1) \\ \vdots & \ddots & \vdots \\ g_1(x_k) & \cdots & g_m(x_k) \end{pmatrix}$$

Зададим линейное преобразование, которое не обязательно в точности восстанавливает исходное признаковое пространство:

$$\widehat{f}_j(x) = \sum_{s=1}^{m} g_s(x) u_{js}, \;\; j = \overline{1, \dots, k}$$

Введем обозначение:

$$U = \big(u_{js}\big)_{n \times m}$$

Необходимо, чтобы восстановленное признаковое пространство, как можно меньше отличалось от исходного. Для этого необходимо решить задачу минимизации:

$$\|GU^T - F\|^2 \to \min_{G,U}.$$

Для достижения минимума необходимо, чтобы U и G были ортогональны, а столбцы матрицы U являлись нормированными собственными векторами матрицы $F^T F$, соответствующими m максимальным собственным числам, при этом:

$$G = FU$$

Данные собственные векторы и называются главными компонентами.

Дисперсия, которую объясняет каждая главная компонента, равна собственному числу λ_i матрицы $F^T F$, которое соответствует данной компоненте.

Долю, объясненной m главными компонентами, дисперсии можно найти по формуле:

$$\frac{\sum_{i=1}^{m} \lambda_i}{\sum_{i=1}^{n} \lambda_i}$$

2.3 Проверка статистических гипотез

При обнаружении различий между несколькими выборками, необходимо знать, в чем именно состоит это различие. Для этого необходимо сравнить группы попарно. Однако в этом случае возникает проблема множественного сравнения, из-за которой можно сделать неверные выводы.

Непараметрическим методом, учитывающим множественные сравнения является критерий Данна [27].

Для проверки пары выборок по критерию Данна, необходимо провести z-тест для следующей статистики [28]:

$$z = \frac{\overline{W_A} - \overline{W_B}}{\sqrt{\frac{N(N+1)}{12}\left(\frac{1}{n_A} + \frac{1}{n_B}\right)}},$$

где $\overline{W_A}$ и $\overline{W_B}$ – средние ранги двух сравниваемых выборок, n_A и n_B – объемы этих выборок, N – общий объем всех выборок

Глава 3. Анализ данных

Для каждого вида обследования, была произведена стандартизация данных, т.е. данные были преобразованы так, чтобы среднее равнялось 0, а стандартное отклонение 1. Выбор количества главных компонент осуществлялся таким образом, чтобы они объясняли не менее 80% дисперсии.

Для каждого эксперимента представлен лучший полученный результат, показано визуальное представление кластеров в пространстве первых трех главных компонент, а также распределение в этом пространстве предположительных меток заболеваний пациентов.

Также представлены результаты проверки статистически значимых различий между кластерами, основываясь на показателях крови. В таблицах показаны значения p-value критерия Данна попарно для каждых из кластеров. В случае наличия различия по показателю крови (при $p\text{-}value < 0.05$) ячейка выделена цветом.

3.1 Выбор инструментария

Для проведения исследования был выбран язык Python 3.7, включающий все необходимые библиотеки и позволяющий работать с большим количеством данных.

Были использованы следующие библиотеки:

- Для извлечения и сбора данных: Pandas;
- Для реализации алгоритмов кластеризации и метода главных компонент: Scikit-learn и Pyclustering;
- Для проведения статистических тестов: scikit-posthocs;
- Для построения графиков: Matplotlib.

3.2 Данные теста Векслера

При кластеризации пациентов по данным результатов теста Векслера наилучший результат показала агломеративная кластеризация с расстоянием

Уорда. Было выделено четыре кластера (рис. 7, рис. 8).

В первом и втором кластерах преобладают пациенты с болезнью Альцгеймера и сосудистой деменцией и практически отсутствуют пациенты из контрольной группы. В третьем кластере наоборот преобладают пациенты с депрессией и из контрольной группы (рис. 9). Однако в четвертом кластере все заболевания представлены в равной степени.

Кроме того, первый и третий кластеры имеют различия в показателях «Кортизол» и «ГП в эритроцитах», а второй и третий кластеры различаются по показателям крови «ГП в эритроцитах» и «Каталаза» (табл. 2).

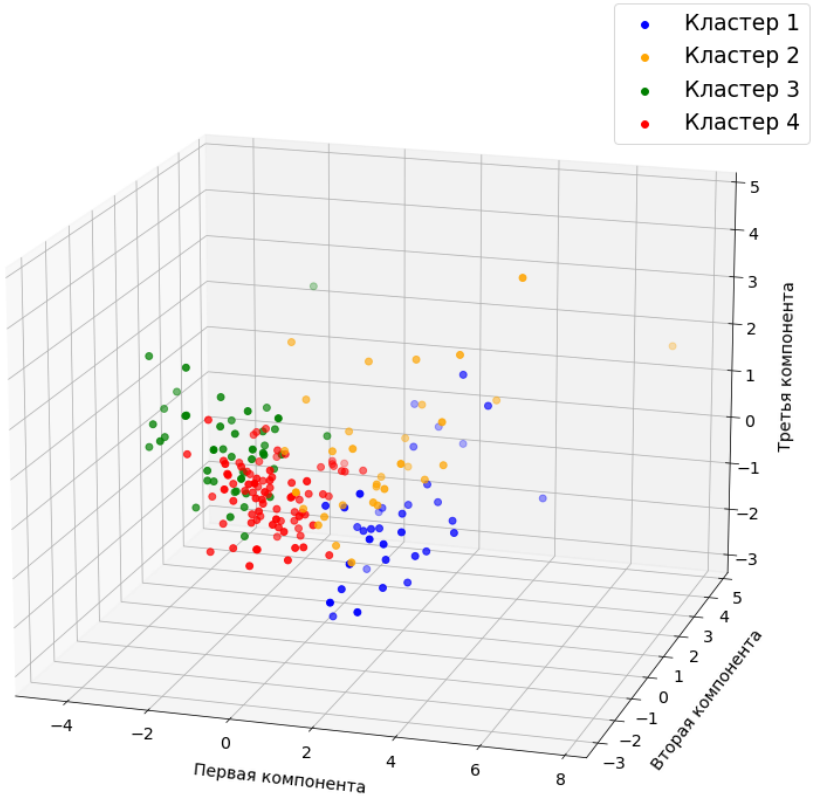

Рис.7. - Результат агломеративной кластеризации по данным теста Векслера

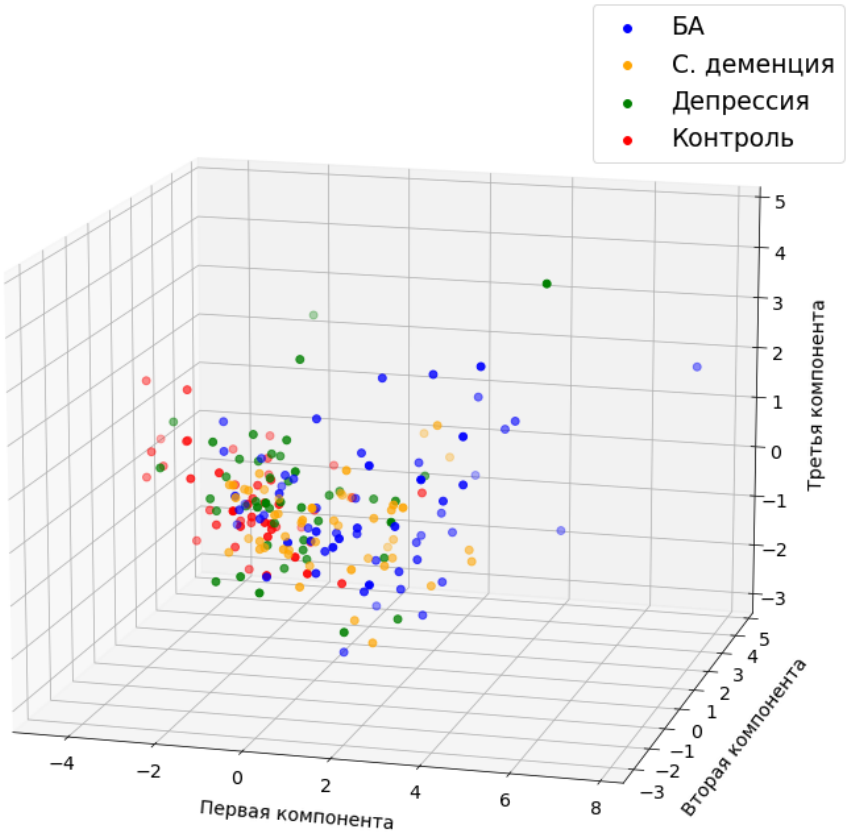

Рис. 8. - Разбиение пациентов с метками предположительных диагнозов (тест Векслера)

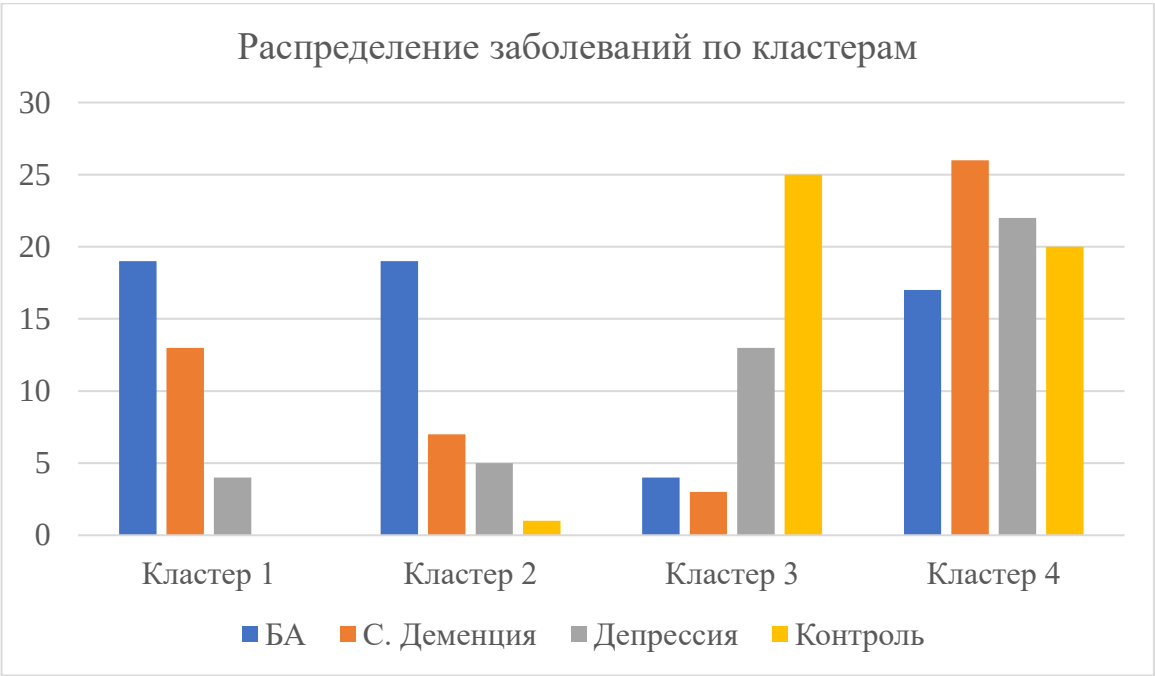

Рис. 9. - Распределение заболеваний по кластерам по данным теста Векслера

Показатель\Кластеры	**1-2**	**1-3**	**1-4**	**2-3**	**2-4**	**3-4**
Цитогенетика	0,999	0,999	0,999	0,999	0,999	0,999
Альбумин	0,821	0,760	0,305	0,760	0,268	0,760
Холестерин	0,679	0,900	0,689	0,526	0,900	0,526
Общий белок	0,931	0,201	0,401	0,252	0,428	0,657
Глюкоза	0,958	0,958	0,461	0,955	0,368	0,478
ЛПВП	0,061	0,822	0,748	0,150	0,150	0,822
ЛПНП	0,958	0,958	0,958	0,877	0,966	0,759
Триглицериды	0,928	0,965	0,965	0,803	0,781	0,990
С-реактивный белок	0,870	0,803	0,870	0,544	0,794	0,870
Коэффициент атерогенности	0,929	0,929	0,929	0,929	0,613	0,929
Пролактин	0,998	0,998	0,998	0,992	0,998	0,992
Кортизол	0,498	**0,034**	0,428	0,394	0,929	0,281
Hb	0,963	0,963	0,882	0,963	0,882	0,963
Эритроциты	0,985	0,988	0,988	0,985	0,985	0,988
Тромбоциты	0,646	0,646	0,646	0,999	0,999	0,999
Лейкоциты	0,972	0,894	0,972	0,972	0,994	0,972
СОЭ	0,620	0,274	0,081	0,826	0,635	0,826
Ретикулоциты	0,435	0,090	0,514	0,664	0,664	0,389
СОД Е/мл	0,875	0,635	0,861	0,635	0,861	0,861
СОД Е/г Hb	0,971	0,895	0,971	0,823	0,958	0,895
ГР в эритроцитах Е/г Hb	0,873	0,873	0,854	0,998	0,998	0,998
ГР в плазме Е/л	0,980	0,985	0,985	0,971	0,980	0,985
ГП в эритроцитах Е/л	0,969	**0,012**	0,195	**0,012**	0,195	0,195
Каталаза	0,686	0,686	0,903	**0,018**	0,442	0,158
Тиоловый статус	0,999	0,989	0,999	0,989	0,999	0,989

Таблица 2. - Значение p-value критерия Данна при попарном сравнении кластеров.

3.3 Данные МРТ

Т.к. показателей МРТ очень много, было принято решение проводить кластеризацию отдельно по каждой доле мозга.

При кластеризации по большинству долей мозга в основном получался один большой кластер с несколькими выбросами. Однако, при рассмотрении некоторых долей удалось получить потенциально полезный результат. Наилучшее разбиение было получено при рассмотрении височной и теменной долей одновременно и использовании агломеративной кластеризации с расстоянием Уорда. Были выделены четыре кластера (рис. 10, рис. 11).

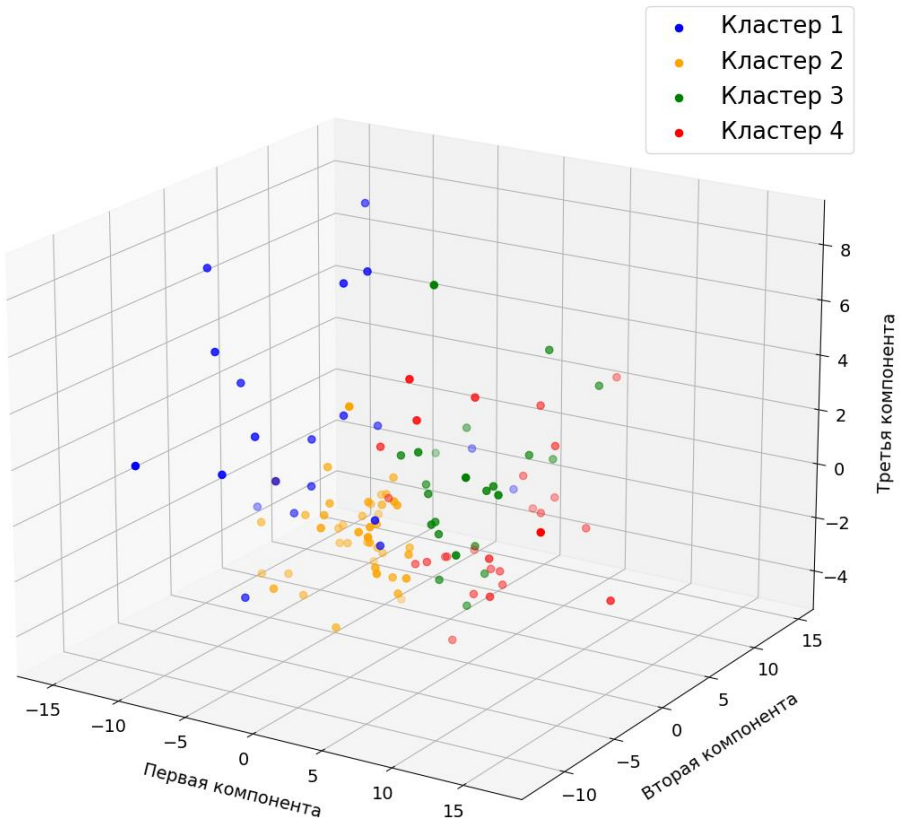

Рис. 10. - Результат агломеративной кластеризации по данным МРТ

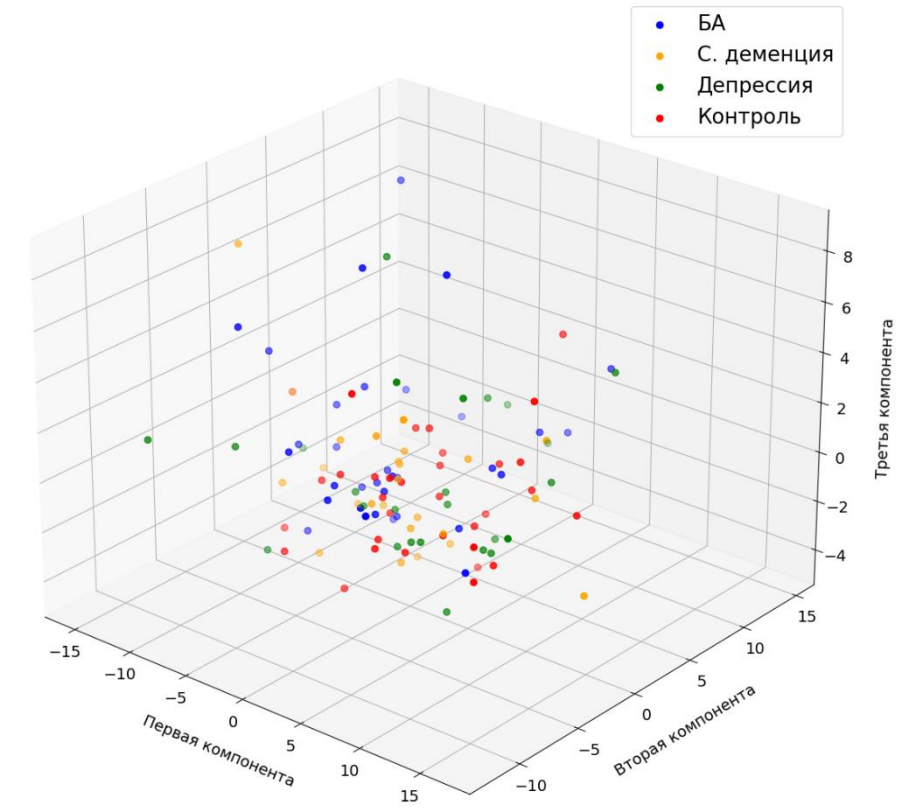

Рис. 11. - Разбиение пациентов с метками предположительных диагнозов (МРТ)

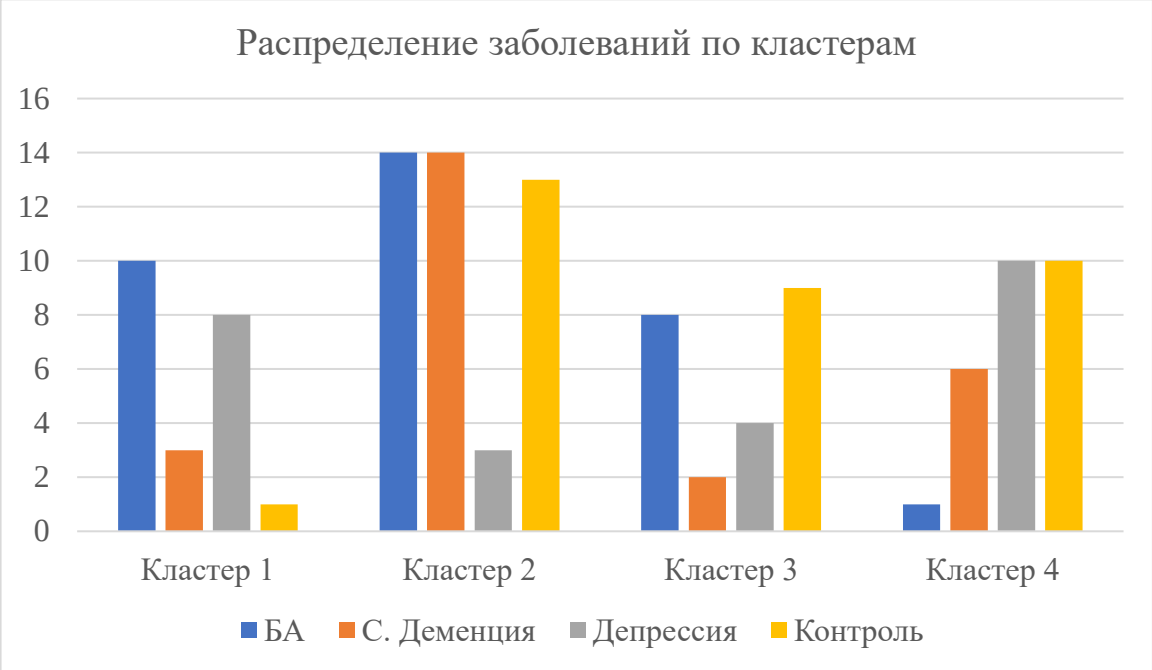

Рис. 12. - Распределение заболеваний по кластерам по данным МРТ

Показатель\Кластеры	**1-2**	**1-3**	**1-4**	**2-3**	**2-4**	**3-4**
Цитогенетика	0,393	0,908	0,479	0,248	0,908	0,381
Альбумин	0,767	0,478	0,867	0,767	0,767	0,478
Холестерин	0,235	0,855	0,889	0,617	0,317	0,889
Общий белок	0,962	0,962	0,962	0,962	0,962	0,962
Глюкоза	0,976	0,976	0,974	0,974	0,951	0,976
ЛПВП	0,649	0,484	0,781	0,781	0,769	0,649
ЛПНП	0,610	0,682	0,780	0,088	0,407	0,696
Триглицериды	0,958	0,994	0,958	0,958	0,994	0,958
С-реактивный белок	0,851	0,932	0,851	0,981	0,981	0,981
Коэффициент атерогенности	1,000	0,792	1,000	0,792	1,000	0,792
Пролактин	0,985	0,870	0,870	0,870	0,870	0,985
Кортизол	0,488	0,644	0,457	0,965	0,965	0,965
Hb	0,510	0,850	0,850	0,783	0,266	0,803
Эритроциты	0,973	0,986	0,986	0,983	0,983	0,986
Тромбоциты	0,688	0,977	0,993	0,755	0,579	0,977
Лейкоциты	0,973	0,937	0,973	0,943	0,973	0,973
СОЭ	0,801	0,493	0,801	0,801	0,801	0,801
Ретикулоциты	0,483	0,992	0,483	0,483	0,992	0,483
СОД Е/мл	0,315	0,341	0,034	0,969	0,407	0,407
СОД Е/г Hb	0,254	0,725	0,341	0,725	0,985	0,725
ГР в эритроцитах Е/г Hb	0,403	0,189	0,864	0,570	0,123	0,031
ГР в плазме Е/л	0,950	0,950	0,950	0,965	0,793	0,802
ГП в эритроцитах Е/л	0,700	0,700	0,617	0,966	0,966	0,966
Каталаза	0,660	0,660	0,660	0,999	0,999	0,999
Тиоловый статус	0,238	0,851	0,878	0,878	0,446	0,878

Таблица 3. - Значимость различий между кластерами по критерию Данна.

В первом кластере практически нет пациентов из контрольной группы, а в четвертом практически отсутствуют пациенты с болезнью Альцгеймера, в двух других кластерах не было заметного преобладания определенной группы (рис. 12). Так же первый и четвертый кластер имеют значимое различие по показателю крови: «СОД Е/мл», а второй и четвертый кластеры по показателю «ГР в эритроцитах» (табл. 3).

3.4 Данные ЭЭГ

При кластеризации по данным ЭЭГ наилучший результат показал агломеративный метод со средним расстоянием, были выделены 4 кластера. (рис. 13, рис. 14).

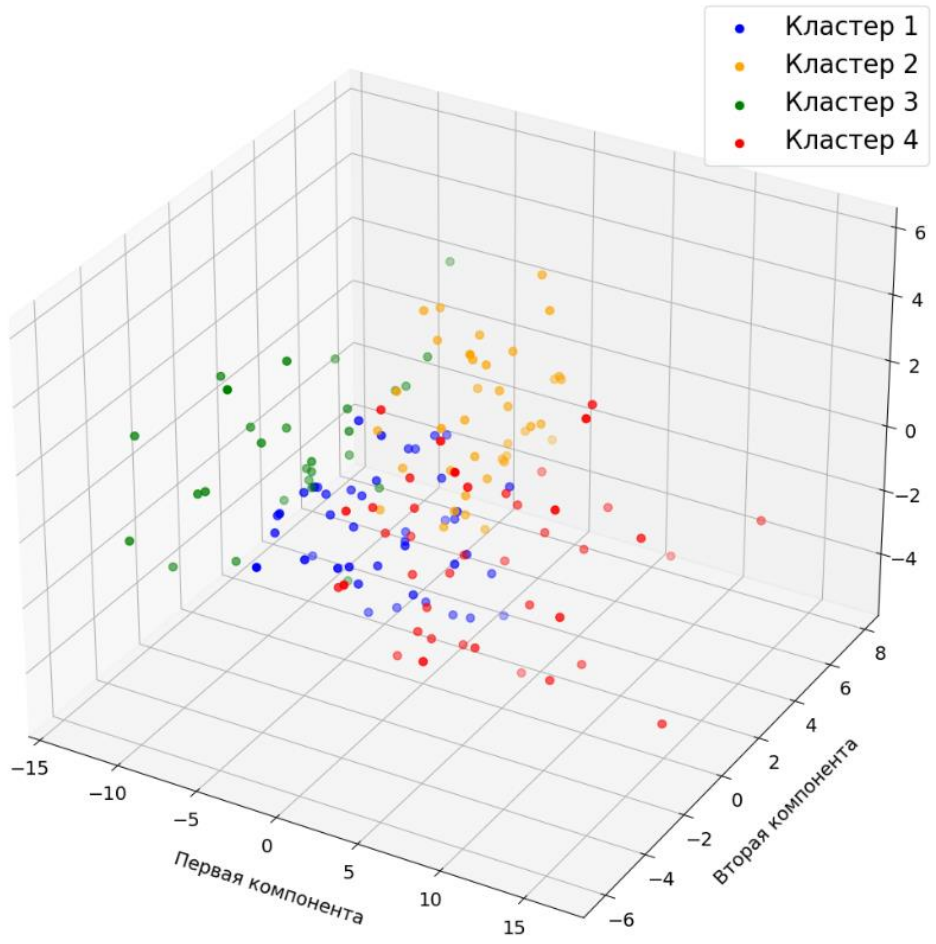

Рис. 13. - Результат агломеративной кластеризации по данным ЭЭГ

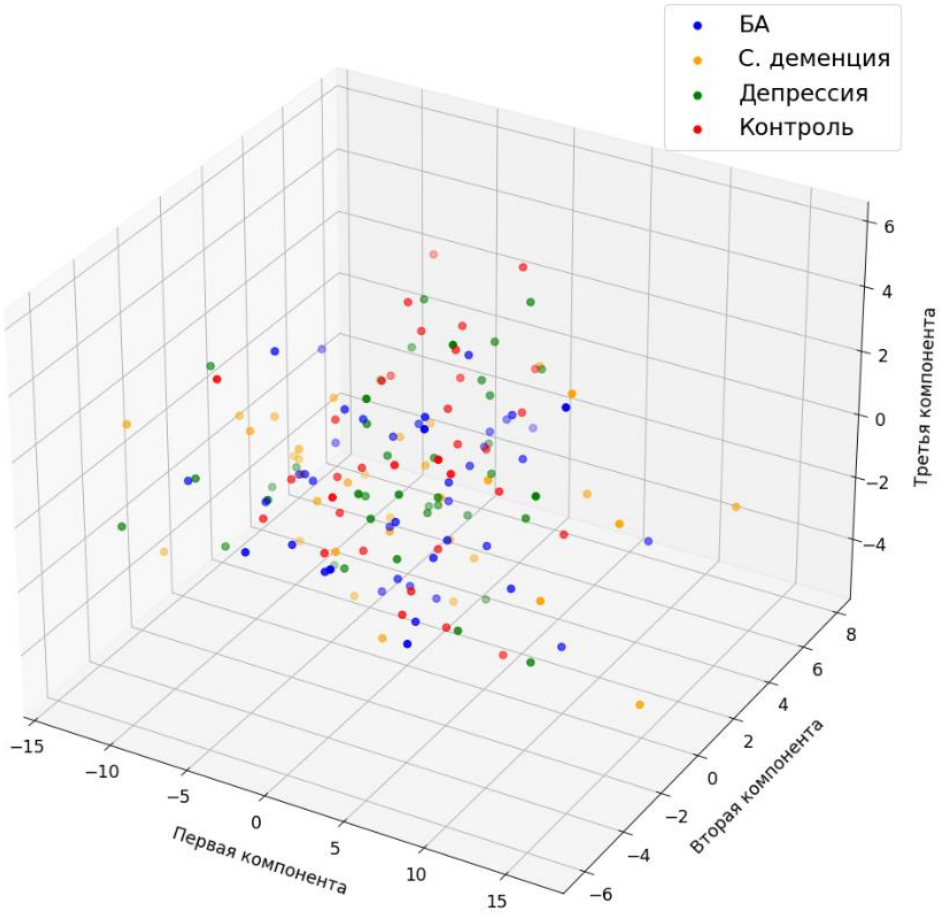

Рис. 14. - Разбиение пациентов с метками предположительных диагнозов (ЭЭГ)

Первый, третий и четвертый кластеры представляют смесь различных групп. Во втором кластере практически отсутствуют пациенты с деменцией (рис. 15). Были найдены статистически значимые различия по анализам крови между третьим и четвертым кластерами по показателям: «Эритроциты» и «СОЭ»; между вторым и третьим кластерами по показателю «Лейкоциты»; между первым и четвертым кластером по показателю «СОЭ»; между первым и третьим по показателю «Тиоловый статус» (табл. 4).

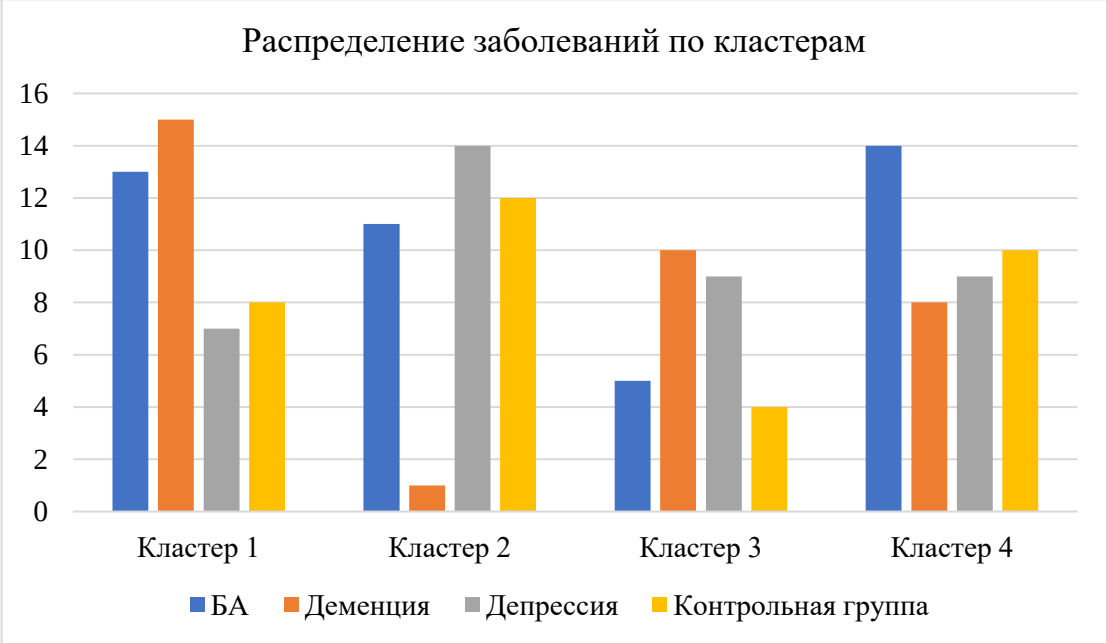

Рис. 15. - Распределение заболеваний по кластерам по данным ЭЭГ

Показатель\Кластеры	**1-2**	**1-3**	**1-4**	**2-3**	**2-4**	**3-4**
Цитогенетика	0,756	0,777	0,925	0,925	0,649	0,756
Альбумин	0,925	0,955	0,738	0,955	0,955	0,925
Холестерин	0,982	0,945	0,982	0,979	0,982	0,979
Общий белок	0,895	0,629	0,895	0,895	0,895	0,598
Глюкоза	0,996	1,000	1,000	0,996	0,996	1,000
ЛПВП	0,750	0,750	0,986	0,257	0,750	0,750
ЛПНП	0,951	0,977	0,977	0,977	0,957	0,977
Триглицериды	0,623	0,482	0,790	0,125	0,598	0,580
С-реактивный белок	0,963	0,950	0,963	0,902	0,937	0,963
Коэффициент атерогенности	0,908	0,854	0,908	0,640	0,908	0,908
Пролактин	0,105	0,219	0,219	0,977	0,964	0,977
Кортизол	0,446	0,997	0,438	0,489	0,997	0,489
Hb	0,650	0,469	0,283	0,459	0,459	0,057
Эритроциты	0,694	0,286	0,360	0,360	0,286	0,014
Тромбоциты	0,873	0,674	0,674	0,890	0,890	0,890
Лейкоциты	0,113	0,605	0,796	0,032	0,143	0,595
СОЭ	0,445	0,628	0,044	0,367	0,445	0,044
Ретикулоциты	0,981	0,857	0,981	0,903	0,981	0,857
СОД Е/мл	0,978	0,978	0,978	0,978	0,864	0,978
СОД Е/г Hb	0,933	0,933	0,599	0,933	0,534	0,534
ГР в эритроцитах Е/г Hb	0,937	0,937	0,967	0,967	0,890	0,937
ГР в плазме Е/л	0,146	0,263	0,064	0,979	0,979	0,979
ГП в эритроцитах Е/л	0,993	0,983	0,993	0,969	0,993	0,983
Каталаза	0,175	0,175	0,900	0,900	0,175	0,175
Тиоловый статус	0,081	0,015	0,255	0,384	0,466	0,255

Таблица 4. - Значимость различий между кластерами по критерию Данна.

3.5 Отбор признаков

Исходя из предположения, что тест Векслера может служить индикатором изменений в некоторых областях мозга, были выбраны показатели МРТ и ЭЭГ, наиболее сильно коррелирующие с результатами теста Векслера.

Из этих признаков была выбрана комбинация, которая дает наиболее точную кластеризацию. Кроме этого, к отобранным признакам по одному добавлялись новые признаки из всех видов обследований таким образом, что на каждой итерации качество кластеризации возрастало. Однако закономерностей в отобранных признаках обнаружено не было.

Лучший результат показал алгоритм k-средних, были выделены 3 кластера (рис. 16, рис. 17).

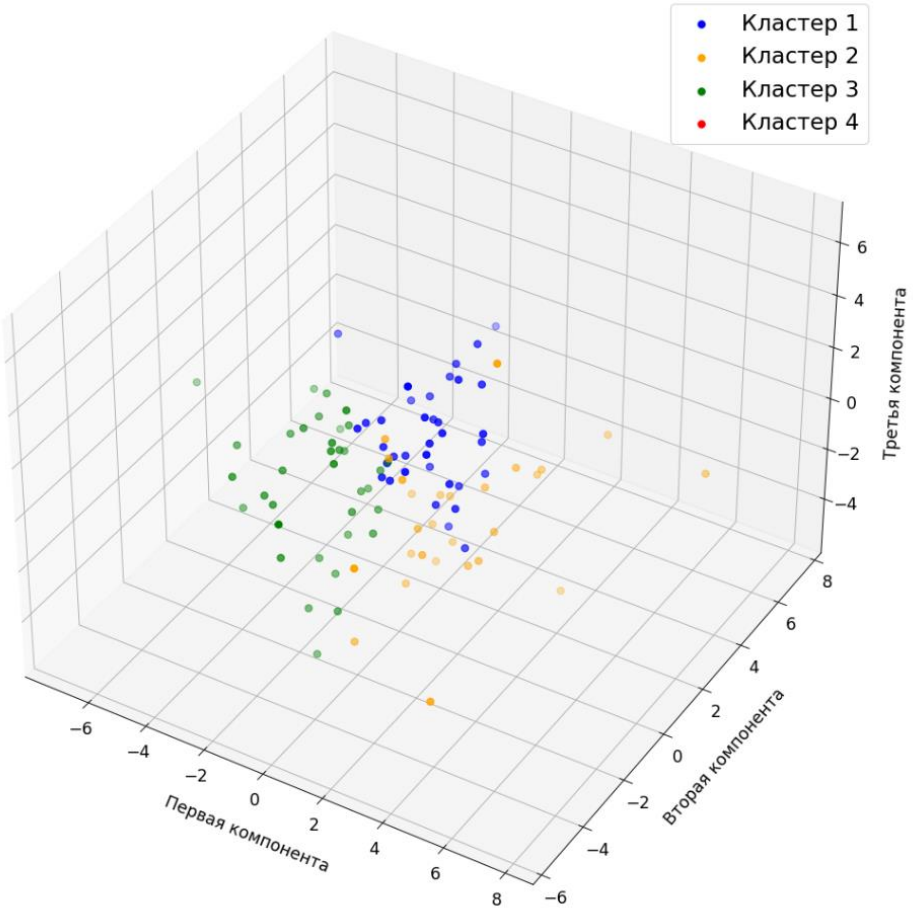

Рис. 16. - Результат иерархической кластеризации по отобранным признакам

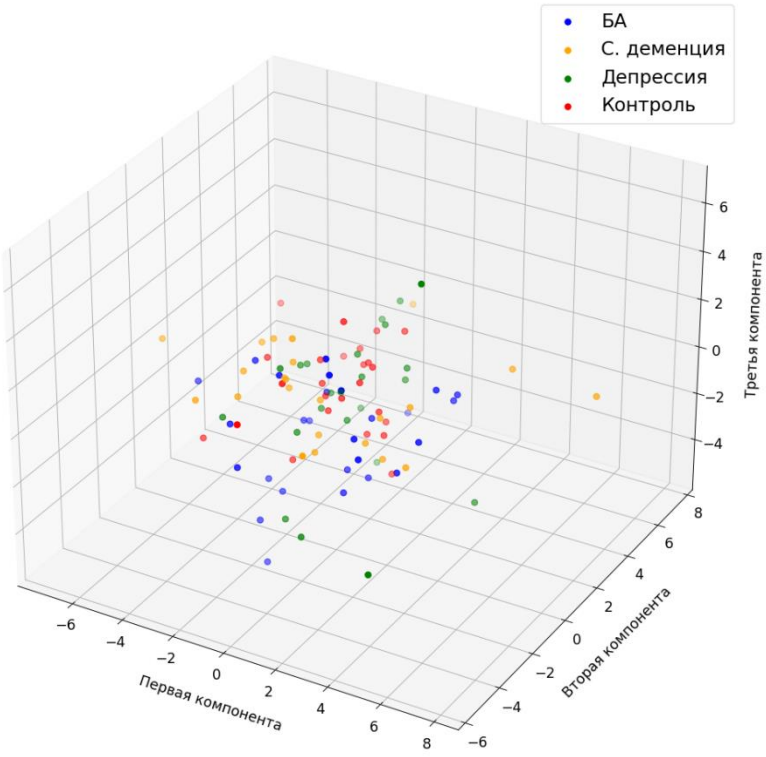

Рис. 17. - Разбиение пациентов с метками предположительных диагнозов (отобранные признаки)

В первом кластере практически отсутствовали пациенты с болезнью Альцгеймера и сосудистой деменцией, во втором кластере отсутствовали пациенты из контрольной группы (рис. 18).

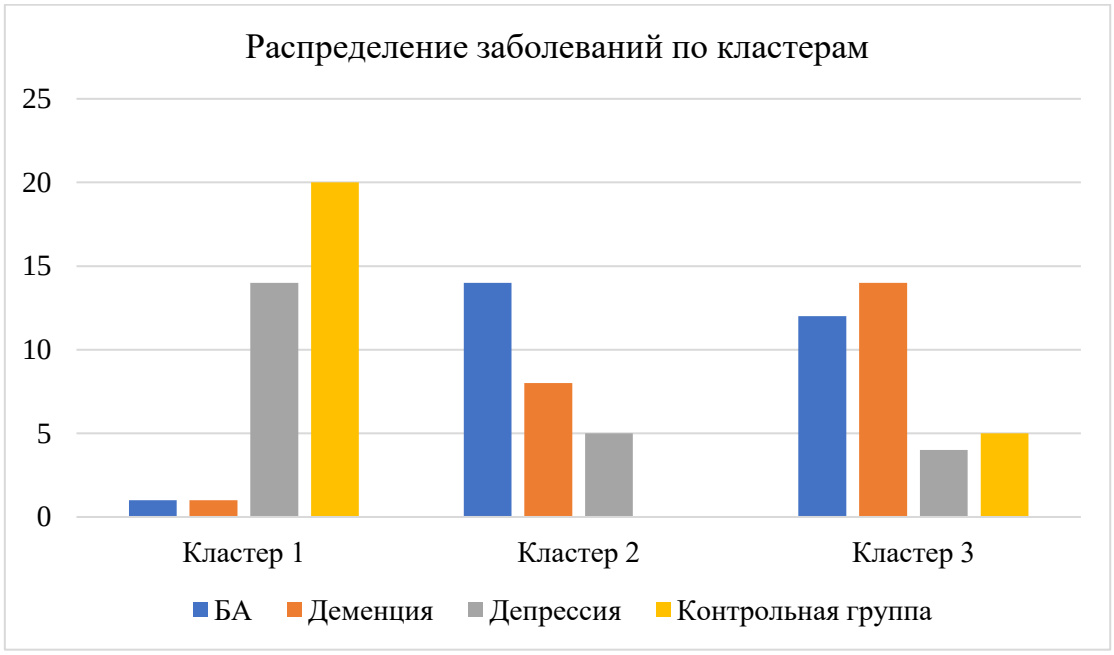

Рис. 18. - Распределение заболеваний по кластерам по отобранным признакам

Кроме того, первый и второй кластеры имеют различия по показателям крови: «Кортизол», «СОД Е/мл», «СОД Е/г Hb», «Каталаза», а второй и третий кластеры по показателям: «Hb», «Эритроциты», «СОД Е/г Hb» (табл. 5).

Показатель\Кластеры	**1-2**	**1-3**	**2-3**
Цитогенетика	0,775	0,964	0,775
Альбумин	0,577	0,698	0,674
Холестерин	0,423	0,290	0,759
Общий белок	0,389	0,551	0,587
Глюкоза	0,960	0,922	0,922
ЛПВП	0,961	0,961	0,961
ЛПНП	0,271	0,271	0,859
Триглицериды	0,898	0,774	0,898
С-реактивный белок	0,233	0,417	0,521
Коэффициент атерогенности	0,941	0,895	0,941
Пролактин	0,951	0,342	0,342
Кортизол	0,010	0,179	0,179
Hb	0,138	0,082	0,002
Эритроциты	0,261	0,261	0,026
Тромбоциты	0,716	0,368	0,716
Лейкоциты	0,937	0,937	0,937
СОЭ	0,442	0,325	0,134
Ретикулоциты	0,991	0,991	0,991
СОД Е/мл	0,010	0,161	0,201
СОД Е/г Hb	0,008	0,791	0,012
ГР в эритроцитах Е/г Hb	0,731	0,865	0,731
ГР в плазме Е/л	0,971	0,971	0,971
ГП в эритроцитах Е/л	0,145	0,476	0,476
Каталаза	0,027	0,098	0,381
Тиоловый статус	0,838	0,838	0,838

Таблица 5. - Значимость различий между кластерами по критерию Данна.

Выводы

Таким образом, проведен анализ имеющихся данных. Было получено, что пациентов не удается разделить так, чтобы все кластеры имели значительные различия. Однако, в некоторых случаях удавалось выделить один-два кластера, которые значимо отличались от остальных по показателям крови (которые, согласно различным исследованиям, могут изменяться при тех или иных заболеваниях мозга), а также содержали или, наоборот, не содержали пациентов с каким-то конкретным предположительным диагнозом.

Также стоит отметить, что наиболее качественную кластеризацию при использовании только одного источника данных удалось получить на данных результатов теста Векслера. Кроме того, при рассмотрении признаков, коррелирующих с его результатами, удалось повысить качество кластеризации. В то же время, только на данных ЭЭГ обследований не удалось получить четких результатов. Однако, добавление ЭЭГ признаков к признакам других обследований, позволило улучшить результаты. В частности, наибольший вклад вносили значения когерентности в альфа-ритме, который согласно мнению врачей, может быть наиболее информативным при диагностике заболеваний мозга.

Наиболее перспективным вариантом получения новых результатов видится увеличение количества пациентов. Кроме того, специалистами из НИПНИ им. Бехтерева был отмечен высокий потенциал продолжения исследования.

Заключение

Все поставленные задачи данной работы были выполнены. Были произведены сбор и обработка данных, необходимых для проведения исследования, в том числе когерентный анализ ЭЭГ. Рассмотрены и применены различные методы кластерного анализа. Выполнен поиск статистически значимых различий между полученными кластерами. Наиболее качественные результаты были получены на признаках, которые наиболее сильно коррелируют с результатами теста Векслера.

Список литературы

1. Patterson C. World alzheimer report 2018. London: Alzheimer's Disease International, 2018. 48 p.
2. What Is Alzheimer’s Disease? - Alzheimer's Association [Электронный ресурс] //URL: https://www.alz.org/alzheimers-dementia/what-is-alzheimers (дата обращения 21.04.2021)
3. Болезнь Альцгеймера и деменция в России - Alzheimer's Association [Электронный ресурс] //URL: https://www.alz.org/ru/деменция-болезнь_Альцгеймера-Россия.asp (дата обращения 21.04.2021)
4. Ridgway G. R. et al. Early-onset Alzheimer disease clinical variants: multivariate analyses of cortical thickness //Neurology. – 2012. – Т. 79. – №. 1. – С. 80-84.
5. Sørensen L. et al. Differential diagnosis of mild cognitive impairment and Alzheimer's disease using structural MRI cortical thickness, hippocampal shape, hippocampal texture, and volumetry //NeuroImage: Clinical. – 2017. – Т. 13. – С. 470-482.
6. Gupta Y. et al. Alzheimer’s disease diagnosis based on cortical and subcortical features //Journal of healthcare engineering. – 2019. – Т. 2019.
7. Bindhi M., Chavez K., Ristanto T. Classification of Alzheimer’s Disease using Patients’ MRI and Related Features // CS229 Final report. – 2017.
8. Lebedeva A. K. et al. MRI-based classification models in prediction of mild cognitive impairment and dementia in late-life depression //Frontiers in aging neuroscience. – 2017. – Т. 9. – С. 13.
9. Trambaiolli L. R. et al. Improving Alzheimer's disease diagnosis with machine learning techniques //Clinical EEG and neuroscience. – 2011. – Т. 42. – №. 3. – С. 160-165.
10. Baker M. et al. EEG patterns in mild cognitive impairment (MCI) patients //The open neuroimaging journal. – 2008. – Т. 2. – С. 52.

11. Gomar J. J. et al. Utility of combinations of biomarkers, cognitive markers, and risk factors to predict conversion from mild cognitive impairment to Alzheimer disease in patients in the Alzheimer's disease neuroimaging initiative //Archives of general psychiatry. – 2011. – Т. 68. – №. 9. – С. 961-969.
12. Butters N. et al. Differentiation of amnesic and demented patients with the Wechsler Memory Scale-Revised //The Clinical Neuropsychologist. – 1988. – Т. 2. – №. 2. – С. 133-148.
13. Seelye A. M. et al. Wechsler Memory Scale–III Faces test performance in patients with mild cognitive impairment and mild Alzheimer's disease //Journal of clinical and experimental neuropsychology. – 2009. – Т. 31. – №. 6. – С. 682-688.
14. Wang X. et al. Oxidative stress and mitochondrial dysfunction in Alzheimer's disease //Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease. – 2014. – Т. 1842. – №. 8. – С. 1240-1247.
15. Незнанов Н. Г. и др. Исследование параметров окислительного стресса при психических нарушениях в позднем возрасте (болезнь Альцгеймера, сосудистая деменция, депрессивное расстройство) //Обозрение психиатрии и медицинской психологии им. ВМ Бехтерева. – 2013. – Т. 4. – С. 31-8.
16. Кулаичев А. П. Метрология вычислительного анализа ЭЭГ //Актуальные проблемы гуманитарных и естественных наук. – 2018. – №. 8. – С. 17-22.
17. Siwek K. et al. Analysis of medical data using dimensionality reduction techniques //Przegląd Elektrotechniczny. – 2013. – Т. 89. – №. 2a. – С. 279--281.
18. Мандель Б. Р. Психогенетика. Иллюстрированное учебное пособие. – Directmedia, 2014.

19. FreeSurfer — open source software suite for processing and analyzing brain MRI images [Электронный ресурс] // URL: http://surfer.nmr.mgh.harvard.edu/ (дата обращения: 21.04.2021)
20. Мельникова Т. С., Лапин И. А., Саркисян В. В. Обзор использования когерентного анализа ЭЭГ в психиатрии //Социальная и клиническая психиатрия. – 2009. – Т. 19. – №. 1.
21. Галанин И. В. и др. Современное состояние проблемы нейропластичности в психиатрии и неврологии //Вестник северо-западного государственного медицинского университета им. ИИ Мечникова. – 2015. – Т. 7. – №. 1.
22. Воронцов К. В. Математические методы обучения по прецедентам (теория обучения машин) //Москва. – 2011. – С. 119-121.
23. Шитиков В. К., Мастицкий С. Э. Классификация, регрессия и другие алгоритмы Data Mining с использованием R //ВК Шитиков, СЭ Мастицкий—: Тольятти, Лондон—2017г. – 2017.
24. K Means - Stanford CS221 [Электронный ресурс] // URL: https://stanford.edu/~cpiech/cs221/handouts/kmeans.html (дата обращения: 21.04.2021)
25. Plot Hierarchical Clustering Dendrogram – scikit-learn [Электронный ресурс] // URL: https://scikit-learn.org/stable/auto_examples/cluster/plot_agglomerative_dendrogram.html (дата обращения 23.04.2021)
26. Gaussian mixture models – scikit-learn [Электронный ресурс] //URL: https://scikit-learn.org/stable/modules/mixture.html (дата обращения 16.12.2020)
27. Гланц С. Медико-биологическая статистика. Пер. с англ. М.: Практика, 1998.
28. Dinno A. Nonparametric pairwise multiple comparisons in independent groups using Dunn's test //The Stata Journal. – 2015. – Т. 15. – №. 1. – С. 292-300.