РЕФЕРАТ

Проведены исследования молекулярных маркеров иммуновоспаления в сыворотке пациентов с шизофренией и здоровых лиц. В результате построен и обучен ряд моделей искусственного интеллекта с разными архитектурами: логистическая регрессия, глубокая нейронная сеть и деревья решений. Созданные модели позволяют по значениям маркеров классифицировать тестируемый набор признаков нового пациента, то есть определить его принадлежность к одному из двух классов: пациенты с шизофренией и группа здоровых лиц.

ВВЕДЕНИЕ

Сегодня в отношении потенциального применения нейронных сетей в разных областях медицины существует огромное количество исследований российских и зарубежных авторов. Нейронные сети и деревья решений представляют собой нелинейные системы, позволяющие гораздо лучше классифицировать данные, чем обычно используемые линейные методы. Бурное развитие нейронных сетей, а также их приложений в медицине в последние 5-10 лет является, прежде всего, свидетельством их эффективности в анализе большого числа диагностических и прогностических маркеров. Применение таких подходов в идентификации шизофрении, классификации и прогнозировании симптоматики, течения и нейрокогнитивного дефицита является новым и, безусловно, перспективным направлением. Так, в работе Cortes-Briones JA (2021) в журнале Schizophr Res. (2021 Jun 5) приводится следующее заключение. Несмотря на годы исследований, механизмы, управляющие началом, рецидивом, симптоматикой и лечением шизофрении, остаются неуловимыми. Отсутствие соответствующих аналитических инструментов, позволяющих справиться с неоднородностью и сложностью шизофрении, может быть одной из причин этой ситуации. Именно машинное обучение как раздел искусственного интеллекта, недавно предоставило доступный способ моделирования и анализа сложных многомерных нелинейных систем. Беспрецедентная точность алгоритмов глубокого обучения в задачах классификации и прогнозирования произвела революцию в широком спектре научных областей и быстро проникает в исследования шизофрении. Машинное обучение может стать ценным подспорьем для врачей в прогнозировании, диагностике и лечении шизофрении, особенно в сочетании с принципами байесовской статистики. Кроме того, машинное обучение может стать мощным инструментом для раскрытия механизмов, лежащих в основе шизофрении, благодаря растущему числу методов, предназначенных для улучшения интерпретируемости моделей и причинно-следственных связей.

Нейронная сеть - математическая модель, позволяющая решать повсеместно распространенные задачи классификации, идентификации, предсказания и даже решения систем линейных уравнений гигантской размерности. В состав нейронной сети данной проекта будут включены как категориальные, так и непрерывные переменные. В качестве первого варианта выбрана модель пятислойной нейронной сети прямого распространения. Входной слой сети состоит из нескольких десятков исходных показаний маркеров. Количество нейронов скрытых слоев подобран экспериментальным путем. Выходной слой будет состоять из одного нейрона, значение которого соответствует вероятности иметь заболевание. Нейронная сеть заданной архитектуры обучена, а также протестирована на выборке, не участвующей в обучении, что позволяет определить погрешность диагностики.

Ухудшение качества модели может происходить, когда модель используется на пациентах с гендерным, расовым или социально-экономическим происхождением, которые отличаются от групп пациентов, на которых была обучена модель. Классификаторы кожных поражений, которые были обучены в основном на изображениях одного тона кожи, снижают эффективность при оценке на изображениях разных тонов кожи, которые неадекватно представлены в обучающей базе данных (Zhang, A., 2022). Для решения этой проблемы будет использована методика кросс валидации. Это особенно важно в данном исследовании, когда общее число пациентов и здоровых лиц не превышает четырех сотен.

В работе Пеников А.А. (2019) упоминается, что главная аспект при использовании машинного обучения в медицине – это выбор правильной архитектуры и способа обучения сети. Важную роль играют данные, на которых сеть обучается, и их объём. Эффективное применение нейронных сетей может значительно облегчить и ускорить работу врачей. Также согласно мнению зарубежных исследователей, машинное обучение может стать мощным инструментом для раскрытия механизмов, лежащих в основе шизофрении, благодаря растущему числу методов, предназначенных для улучшения интерпретируемости моделей и причинно-следственных связей и ценным подспорьем для врачей в прогнозировании, диагностике и лечении шизофрении.

Диагностические ошибки являются причиной приблизительно 10 процентов смертей пациентов (Kumba Sennaar). Идеальный метод диагностики должен иметь абсолютную чувствительность и специфичность - во-первых, не пропускать ни одного действительно больного человека и, во-вторых, не ошибаться в случае здоровых людей. Чтобы застраховаться, нужно стараться прежде всего обеспечить стопроцентную чувствительность метода - нельзя пропускать заболевание. Но это оборачивается, как правило, низкой специфичностью метода - у многих людей врачи подозревают заболевания, которыми на самом деле пациенты не страдают. Разработка методики на основе искусственного интеллекта позволяет генерировать независимый диагноз и помогать врачу.

Cortes-Briones JA, Tapia-Rivas NI, D'Souza DC, Estevez PA. Going deep into schizophrenia with artificial intelligence. Schizophr Res. 2021 Jun 5:S0920-9964(21)00179-1. doi: 10.1016/j.schres.2021.05.018.

Пеников А.А., Козина А.В., Белов Ю.С. (2019). Искусственные нейронные сети в диагностике многокритериальных заболеваний. *Материалы Всероссийской научно-технической конференции «Наукоемкие технологии в приборо- и машиностроении и развитие инновационной деятельности в вузе»*. С. 149–151;

Kumba Sennaar. Machine Learning for Medical Diagnostic – Current Applications. URL: https://emerj.com/ai-sector-overviews/machine-learning-medical-diagnostics-4-current-applications/

Zhang, A., Xing, L., Zou, J. *et al.* Shifting machine learning for healthcare from development to deployment and from models to data. *Nat. Biomed. Eng* (2022). https://doi.org/10.1038/s41551-022-00898-y

2 МАТЕРИАЛЫ И МЕТОДЫ

2.1 Характеристика исследуемой выборки

2.2 Материалы исследования

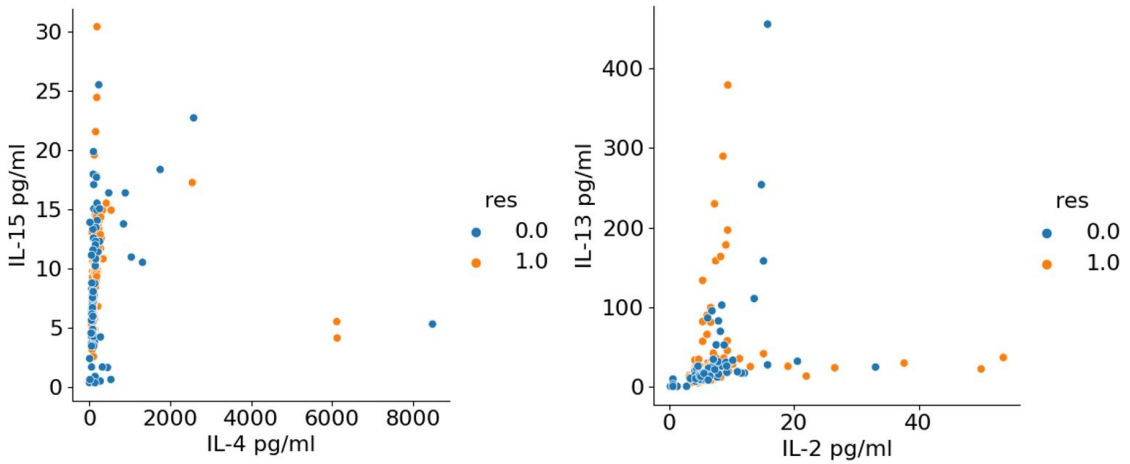
2.3 Методы исследования

Использованные методы основаны на использовании нейронный сетей и машинного обучения. Требуется создать алгоритм, позволяющий находить сложные закономерности и взаимосвязи между различными маркерами при шизофрении. В исследовании реализуется классическая модель логистической регрессии, глубокая нейронная сеть и дерево решений. Распространение и применение искусственного интеллекта и нейросетей неизбежно в медицине и в изучении шизофрении, в частности. Однако остро возникает вопрос контроля уровня достоверности предсказаний нейронной сети в условиях малой обучающей выборки.

2 Статистическая обработка данных

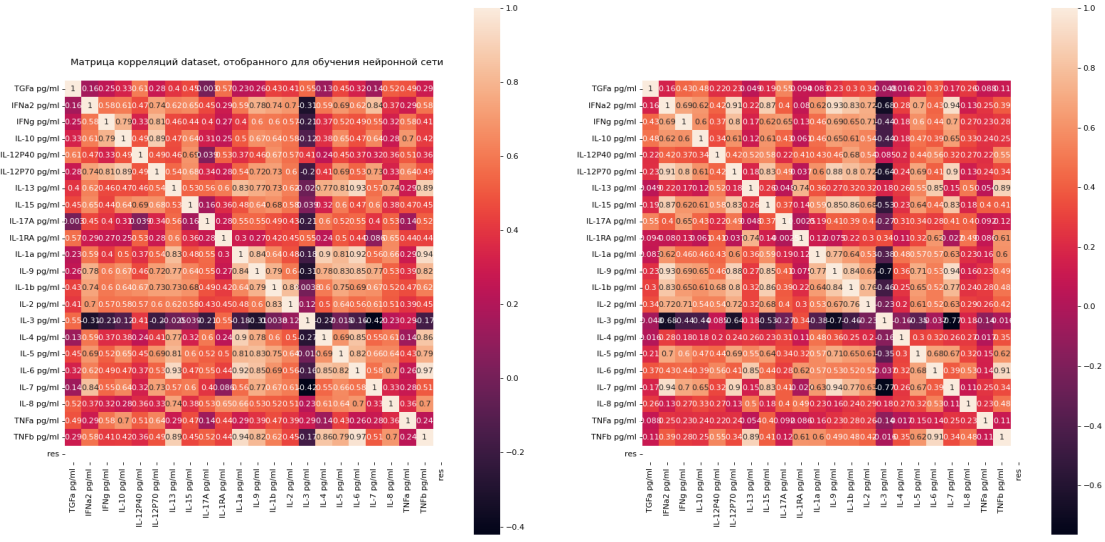
Главная цель ряда задач первого года состояла в исследовании степени классификации (разделяемости) двух классов: пациентов и здоровых лиц, на основе анализа величин молекулярных маркеров иммуновоспаления.

В первую очередь были произведены сравнения набранных спектров следующих маркеров в сыворотке крови пациентов и контрольной группе здоровых лиц: IL-1β, IL-2, IL-3, IL-4, IL-5, IL-6, IL-7, IL-8, IL-9, IL-10, IL-12(p40), IFN-γ, TNF-α, IL-12, IFN-α, IL-1RA, IL-12(p70), IL-13, IL-15, IL-17A, TGF-α, TNF-β. Так, например, на следующем рисунке показано сравнение спектров IL-15 vs IL-4 (слева) и IL-13 vs IL-2 (справа) для двух классов: пациентов (оранжевый цвет) и здоровых лиц (голубой цвет). В результате сравнения не было обнаружено ни одного признака, форма распределения которого существенным образом отличается для пациентов и здоровых лиц.



Сравнение спектров IL-15 vs IL-4 (слева) и IL-13 vs IL-2 (справа) для пациентов (голубой цвет) и здоровых лиц (оранжевый цвет)

Также были сравнены таблицы линейной корреляции маркеров (см. рисунок ниже). Матрицы корреляции имеют ярко выраженный положительный характер и содержат разнообразные значения от -0.7 до 0.95. Видно, что корреляции пациентов и здоровых лиц очень похожи, хотя наибольшее отличие в корреляциях наблюдается в маркерах IL-1b, IL-3, IL-5, IL-7, IL-9, IL-15, IL-12P70, IFNa2, TNFa, TNFb, что вызвано отчасти статистическими выбросами ввиду ограниченной статистики, доступной для анализа.



Матрица линейной корреляции маркеров для здоровых лиц (слева) и пациентов (справа)

Ввиду обстоятельства, что плотность распределения каждого маркера нам неизвестна априори, возникает потребность в эмпирической выборочной плотности и необходимость в проверке статистической устойчивости искомых моделей. Неустойчивость может быть обусловлена как ограниченностью количества событий, так и не статистическими выбросами из-за

неучтенных факторов в процессе измерений. Такая проверка была выполнена для каждой архитектуры, которая будет описана ниже. Для этого используется метод кросс валидации, а именно:

1. Все события обоих классов перемешиваются в случайном порядке.

2. Первые 20% случаев используются для теста, остальные – для тренировки модели.

3. Для следующей модели для теста используются вторые 20%, а остальные – для тренировки и т.д. В результате обучается пять моделей, обученных на перекрывающихся данных.

4. Для каждой модели две выборки: для тренировки и для теста, являются независимыми.

Методика кросс валидации проиллюстрирована рисунком ниже.

Номер выборки	Выборка для тренировки # 264		Выборка для теста # 66	
	# здоровых	# больных	# здоровых	# больных
1	75	189	20	46
2	83	181	12	54
3	73	191	22	44
4	73	191	22	44
5	76	188	19	47

Ниже рассмотрены использованные типы предиктивных моделей.

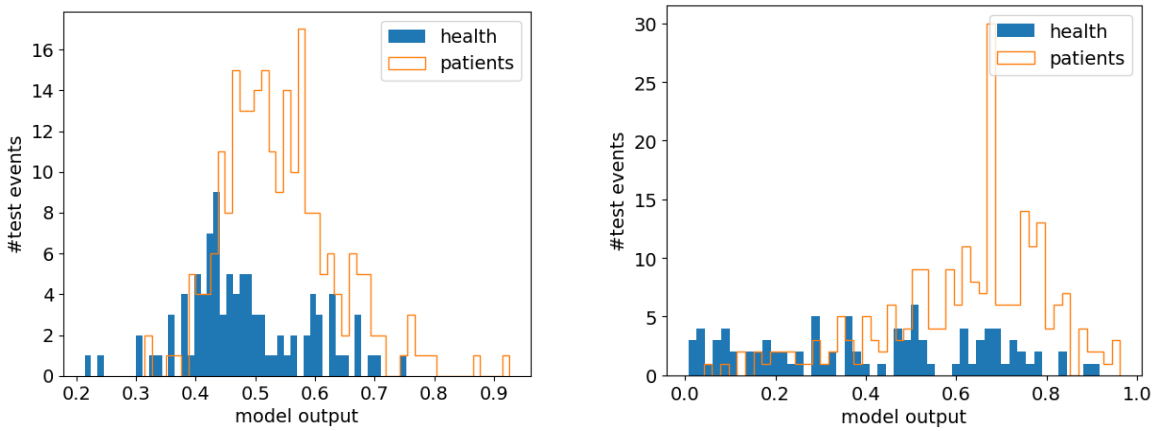
Архитектура №1. Логистическая регрессия. Представляет из себя линейную модель, которую можно представить в виде $f(x) = 1 / [1+ \exp(-z)]$, где $z = \theta 0 + \theta 1x1 + \ldots + \theta nxn$, x – признаки-значения маркеров, θ – неизвестные параметры, определенные в ходе обучения модели. В качестве функции потерь, которая минимизируется в процессе обучения модели, используется функция квадратичной ошибки MSE. Для минимизации применен метод градиентного спуска с оптимизацией методом Adam и penalty функцией в качестве регуляризации. Гиперпараметр learning rate был выбран как 0.0003, количество эпох равно 1500. Такие же настройки обучения использовались для следующей архитектуры.

Архитектура №2. Глубокая нейронная сеть.

Сеть состоит из пяти полносвязных слоев со следующим числом нейронов на каждом слое: 22-22-11-7-1. В качестве функции активации использовалась функция Relu кроме последнего слоя, где вызывалась функция sigmoid. Всего было обучено 851 свободный параметр. Эффективность работы модели по данным тренировочных выборок очень высокая, т.к. параметры модели подбираются таким образом, чтобы разделяемость данных по классам была как можно выше. Однако глубокая нейронная сеть обладает избыточным набором параметров и велика вероятность переобучения (overfitting). Для предотвращения переобучения, оптимизация параметров прекращалась при достижении MSE = 0.15.

Ниже представлены отклики сетей для логистической регрессии (слева) и глубокой нейронной сети (справа). В гистограммах суммированы значения для всех пяти независимых

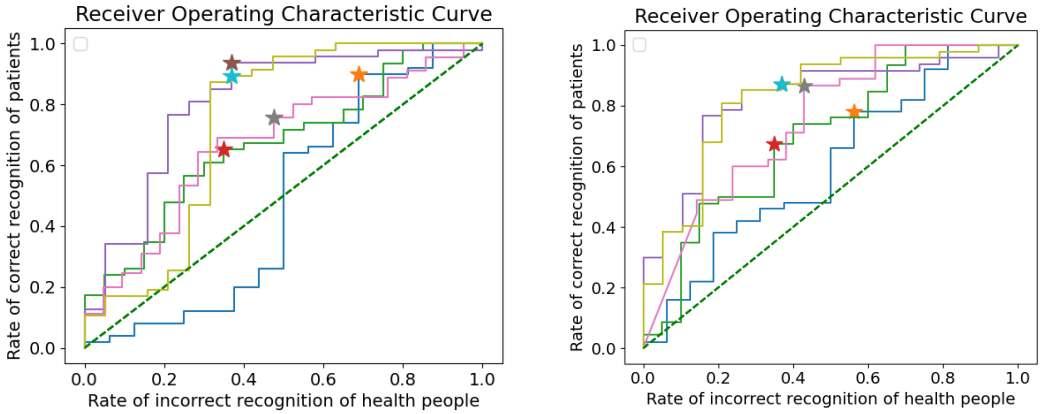
тестовых выборок. Видно, что в качестве критерия классификации пациентов и здоровых лиц нужно использовать значение в районе 0.45: если отклик меньше этого значения, то данные принадлежат классу здоровых, если больше – классу пациентов, больных шизофренией.



Распределения выходных значений логистической регрессии (слева) и глубокой нейронной сети (справа). В гистограммах суммированы значения для всех пяти независимых тестовых выборок

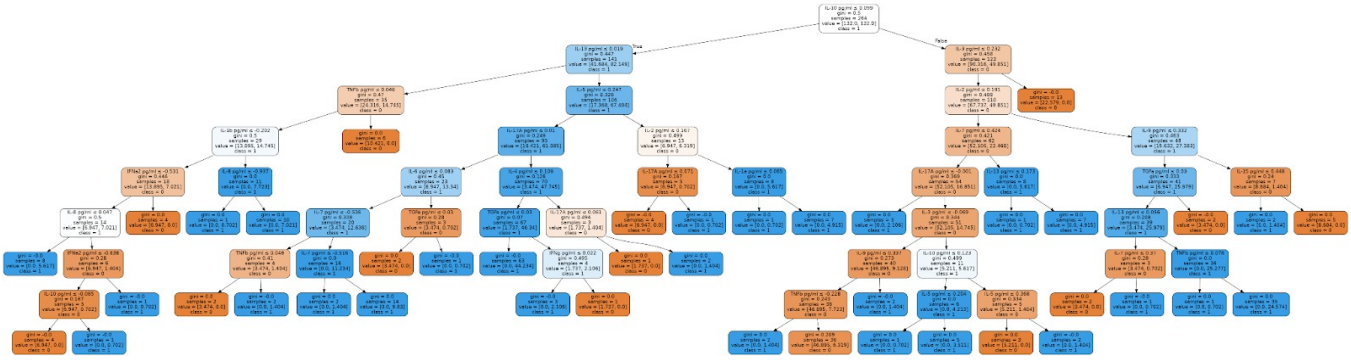
Кроме бинарного анализа возможен вероятностный подход, когда вышепредставленные гистограммы используются в качестве эмпирических плотностей распределений. Тогда, нормированная площадь оранжевой гистограммы левее полученного отклика – это вероятность иметь заболевание, площадь синей гистограммы правее полученного отклика – вероятность принадлежать классу здоровых лиц. Видно, что отклик глубокой нейронной сети больше разнесен влево-вправо и имеет больший потенциал для разделяемости на классы.

На рисунке ниже показаны ROC кривые пяти моделей, проверенных на пяти независимых тестовых выборках, для линейной регрессии (слева) и глубокой нейронной сети (справа). Нужно отдельно подчеркнуть, что проверяемая модель не была знакома в процессе обучения с тестовыми событиями, поэтому это в полной мере моделирует ситуацию анализа данных нового пациента. Звездочками отмечены оптимальные значения (0.45 в обсуждении выше) откликов моделей, которые могут быть использованы для признания принадлежности тому или иному классу. Оптимальные значения соответствуют максимальной вероятности правильного диагноза для пациента, поделенной на минимальную вероятность ложного диагноза для здорового лица. Ввиду ограниченной статистики наблюдается разброс профилей ROC кривых. Диагональная пунктирная линия – это равновероятностная 50/50 классификация.



Receiver operating characteristic (ROC) кривые пяти моделей на пяти независимых тестовых выборках для линейной регрессии (слева) и глубокой нейронной сети (справа)

Архитектура №3. Дерево принятий решений. Каждый внутренняя вершина графа сопоставляется с одним из входных признаков. Пример дерева проиллюстрирован на рисунке ниже. Критерий разделения по каждому признаку в вершинах определяется из требования минимального значения параметра gini, который близок к понятию энтропии – критерия смешанности событий разных классов, т.е. энтропия равна нулю при однозначной 100% разделяемости и наоборот равна единице при полной и равной смешанности событий разных классов. Начальное значение энтропии в исследуемых выборках составляет около 0.86, что определяется тем, что в исследуемом наборе данных число пациентов превышает число здоровых лиц в 2.5 раза, т.е. выборка ассиметрична.



Пример обученного дерева решений классификации с глубиной = 8 и постепенным уменьшением показателя gini

Для предотвращения переобучения модели, максимальная глубина дерева была ограничена как восемь, что позволяет иметь наиболее стабильный результат по итогам кросс валидации. Нужно подчеркнуть, что данная регуляризация в виде ограничения глубины, во-первых, обусловлена ограниченностью исследуемой выборки, во-вторых, лимитирует конечную

эффективность модели. Таким образом, увеличение статистики – это приоритетное усовершенствование, которое может быть предпринято в будущих исследованиях.

В рамках кросс валидации было построено пять деревьев с использованием следующих признаков: IL-2, IL-3, IL-5, IL-7, IL-10, IL-13, TNFa, TNFb, TGFa, а также других признаков, значимость которых меньше, но существенна. Именно кумулятивный эффект от утилизации всех признаков-маркеров позволяет достичь наилучших показателей.

Ниже в таблице приведены сравнения точностей классификации набора маркеров по тестовым выборкам согласно 15 моделям (по пять на три архитектуры, описанные выше). TN / FP (true negative / false positive) означает в данном случае вероятность правильного предсказания для пациента / вероятность ложного предсказания для здоровых лиц. TN должен быть как можно больше, FP – как можно меньше.

	Логистич. регр. TN % / FP %	Глуб. нейр. сеть TN % / FP %	Дерево решений TN % / FP %
Выборка №1	90 / 69	78 / 56	76 / 38
Выборка №2	65 / 35	67 / 35	69 / 60
Выборка №3	94 / 37	87 / 37	83 / 32
Выборка №4	76 / 48	87 / 43	80 / 52
Выборка №5	89 / 37	87 / 37	77 / 53
Средние значения	**83±12 / 45±14**	**81±9 / 42±9**	**77±5 / 47±11**

Стандартные отклонения, указанные в последней строке соответствуют разбросу между тестовыми выборками. Значительные статистические флуктуации обусловлены ограниченностью выборок, доступной для обучения и теста моделей. Видно, что ни одна из моделей не показывает явного преимущества. Глубокая нейронная сеть, которая за счет большого числа внутренних нелинейных связей, способна отследить скрытые закономерности в данных, не демонстрирует свой функционал. Это говорит о том, что либо обучающие выборки данных не являются репрезентативными, либо исследуемые данные не содержат скрытых закономерностей и классификационной мощности линейной регрессии достаточно. Дерево решений – наиболее нативная архитектура для решения задачи классификации, демонстрирует наихудшие показатели, хотя статистические ошибки слишком велики для утвердительного заключения об этом.

Кроме ограниченной статистики, другой важный фактор, лимитирующий показатели моделей машинного обучения, - ассиметричная статистика по количеству пациентов и здоровых лиц. Для учета этой особенности, меньшая выборка здоровых лиц была искусственная расширена с использование методики бутстрап.

3 РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЯ

В результате исследований показано, что анализ на основе искусственного интеллекта молекулярных маркеров иммуновоспаления в сыворотке человека позволяет вычислять вероятностные оценки заболевания пациентом шизофренией. Такие предсказания могут служить как основой для объективного независимого диагноза, так и дополнительным признаком, взятым доктором во внимание.

В результате исследований построен и обучен ряд моделей с разными архитектурами: логистическая регрессия, глубокая нейронная сеть и деревья решений. Созданные модели позволяют по значениям маркеров классифицировать тестируемый набор признаков нового пациента, то есть определить их принадлежность к одному из двух классов: пациенты с шизофренией и группа здоровых лиц. Все три подхода демонстрируют близкие результаты и служат взаимной проверкой своих ответов.

Используя методику независимых тестовых выборок, было показано, что нейронная сеть позволяет идентифицировать по маркерам шизофрению у пациентов с вероятностью 81±9%, и предсказывать ложный диагноз для здоровых лиц с вероятностью 42±9%. Анализ другой линейки маркеров в следующем году проекта может способствовать улучшению работоспособности моделей.

Результаты исследований оформлены в виде программного обеспечения на языке Python, размещены в репозитории github (см. https://github.com/eakozyrev/schizo). Результаты и ход исследований обсуждались на заседании секции "Молекулярная психиатрия".

ЗАКЛЮЧЕНИЕ

Проведенные исследования данных маркеров продемонстрировали свой потенциал в прогнозировании диагноза. Используя методику независимых тестовых выборок, было показано, что модель нейронной сети позволяет идентифицировать по маркерам шизофрению у пациентов с вероятностью 81±9%, и предсказывать ложный диагноз для здоровых лиц с вероятностью 42±9%. Использование новых данных во втором году исследований позволит верифицировать уже полученные результаты, а также увеличить уровень специфичности и чувствительности моделей.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ