ДОКЛАДЫ РОССИЙСКОЙ АКАДЕМИИ НАУК. МАТЕМАТИКА, ИНФОРМАТИКА, ПРОЦЕССЫ УПРАВЛЕНИЯ, 2023, том 514, № 2, с. 242–249

УДК 004.891.3

ДИАГНОСТИКА ТЯЖЕСТИ СИМПТОМОВ ДЕПРЕССИИ ПРИ ПОМОЩИ ОБЪЯСНИМОГО ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

© 2023 г. С. Шалилех[1,2,*], А. О. Копцева[2,], Т. И. Шишковская[3,***], М. В. Худякова[1,4,****], О. В. Драгой[1,5,*****]**

Представлено академиком РАН А.Л. Семеновым
Поступило 01.08.2023 г.
После доработки 18.08.2023 г.
Принято к публикации 15.10.2023 г.

Эта статья представляет исследование, направленное на (i) разработку решения на основе искусственного интеллекта для диагностики депрессии и (ii) изучение психиатрических данных с помощью объяснимого искусственного интеллекта. Авторы собрали и аннотировали новый набор аудиоданных, сформулировали задачу регрессии и изучили производительность восьми ее алгоритмов. Результаты показали, что метод ближайших соседей и случайный лес образуют группу с наиболее приемлемыми результатами. Была определена важность характеристик лучшего алгоритма регрессии и выявлены три наиболее значимые для диагностики характеристики: четвертые коэффициенты мел-частотного кепстрального преобразования, гармоническая разница H1-A1 и хрипота.

Ключевые слова: распознавание депрессии, акустические признаки, регрессия, объяснимый искусственный интеллект

DOI: 10.31857/S268695432360091X, **EDN:** GOCKCK

1. ВВЕДЕНИЕ: ПРЕДПОСЫЛКИ И МОТИВАЦИЯ

Согласно докладу Всемирной Организации Здравоохранения, к первой четверти 2023 г. около 5% процентов взрослых страдают от депрессии. Это одно из самых распространенных психических расстройств, проявляющееся в сниженном эмоциональном фоне и потере интереса к занятиям, которые ранее приносили удовольствие. Тяжелое течение болезни может приводить к суициду – более 700 000 случаев регистрируется ежегодно [1]. Существующие методы диагностики включают опросники, клинические шкалы, интервью и другие формы оценки состояния личности: результаты этих методов зависят от знаний и опыта медицинского сотрудника и подвержены влиянию человеческого фактора. Более того, пациенты с тяжелой формой депрессии могут отказываться от посещения врача, а пациенты со слабо выраженными симптомами могут не подозревать о развивающемся заболевании. По этим причинам крайне важна разработка автоматизированного и доступного диагностического решения, которое будет быстрым и надежным.

Достижения в развитии искусственного интеллекта оказывают влияние на разные аспекты жизни людей. В частности, были разработаны решения, позволяющие выявлять психические заболевания. Выявление депрессии при помощи искусственного интеллекта с использованием данных изображений, видео и аудио представляется естественным решением – пациенты с депрессией демонстрируют отличные от здоровых людей мимические и голосовые проявления. Из трех перечисленных типов данных наиболее предпочтительными являются аудиоданные в силу простоты записи и минимального дискомфорта, доставляемого пациенту.

[1]*Центр языка и мозга, Научно-исследовательский университет "Высшая школа экономики", Москва, Россия*
[2]*Научно-учебная лаборатория моделирования зрительного восприятия и внимания, Научно-исследовательский университет "Высшая школа экономики", Москва, Россия*
[3]*Отделение эндогенных психических расстройств и аффективных состояний ФГБУН "Центр психического здоровья", Москва, Россия*
[4]*Центр языка и мозга, Научно-исследовательский университет "Высшая школа экономики", Нижний Новгород, Россия*
[5]*Институт языкознания, Российская академия наук, Москва, Россия*
**e-mail: sr.shalileh@gmail.com*
***e-mail: akoptseva@hse.ru*
****e-mail: tszyszkowska@gmail.com*
*****e-mail: mariya.kh@gmail.com*
******e-mail: odragoy@hse.ru*

Авторы недавнего обзора исследований в области распознавания депрессии при помощи речи [12] систематизировали существующие к середине 2022 г. методы искусственного интеллекта, наборы данных и возникающие трудности. Они разделили существующие статьи на те, в которых признаки были отобраны вручную, и те, где признаки были сгенерированы при помощи глубинного обучения. В первом случае требуются знания в области психиатрии, второй подход полностью ориентирован на вычисления.

Мы сфокусировались на использовании аудиоданных для разработки метода выявления депрессии, основанного на искуственном интеллекте, с использованием как отобранных вручную, так и сгенерированных методами глубинного обучения признаков. К нашему удивлению, несмотря на все усилия, методы глубинного обучения, такие как извлечение спектрограмм и передача их в сверточные нейронные сети с различными архитектурами, не дали обнадеживающих результатов, поэтому мы сконцентрировались на отобранных вручную признаках.

Несмотря на то что с вычислительной точки зрения данное исследование не содержит существенных нововведений, мы формулируем наш вклад следующим образом:

1. Мы собрали новый уникальный набор аудиоданных на русском языке, который позднее был добавлен в корпус Discourse Diversity Database (3D corpus) [9].

2. Это первое и наиболее обстоятельное исследование применения восьми методов машинного обучения к новым русскоязычным данным.

3. Мы подробно изучили влияние лучшего регрессионного подхода при помощи метода SHAP для проверки существующего психиатрического знания.

4. Мы эмпирически изучили эффективность трех заданий для элиситации связной речи для выявления депрессии на основе аудиоданных.

2. ДАННЫЕ И ПРОЦЕДУРА ВЫЧИСЛЕНИЙ

Участники и данные

Наш уникальный набор данных состоит из аудиоданных 247 участников: 151 здорового участника и 96 участников с депрессивными симптомами. 35 из 96 участников с депрессивными симптомами были пациентами Научного центра психического здоровья (НЦПЗ) в Москве, и их состояние было оценено психиатрами. Остальные заполнили опросники для оценки: (i) широкого спектра психологических и психопатологических проблем, (ii) симптомов мании, (iii) симптомов депрессии. Участники с манией или расстройствами мышления были исключены из исследования. Полученные численные оценки были переведены в шкалу от 0 до 3, где 0 соответствует отсутствию симптомов депрессии (контрольная группа), а 3 соответствует наличию тяжелых симптомов.

После оценки состояния участников были записаны их образцы речи при выполнении трех заданий: (i) рассказ по серии рисунков, (ii) личная история, (iii) инструкция по серии рисунков. Для каждого из заданий было три варианта: один из трех комиксов Херлуфа Бидструпа для рассказа по рисункам, один из трех вопросов о значимых событиях жизни участника (наиболее яркие воспоминания), и инструкции по самостоятельному сбору мебели из Икеа. Порядок выполнения заданий был фиксирован. Более подробное описание приведено в [9].

Никто из участников не сообщил об истории неврологических заболеваний или зависимостей. Всеми участниками было подписано информированное согласие, исследование было одобрено Комиссией по этике НЦПЗ.

Акустические признаки

Мы извлекли набор признаков eGeMAPS [4] при помощи openSMILE [3]. eGeMAPS состоит из 88 параметров, 16 из которых относятся к перцентилям громкости или высоты голоса. Поскольку некоторые из наших аудиофайлов содержат существенный фоновый шум, эти 16 параметров были исключены из вычислений. Обозначим среднее значение сигнала μ, его стандартное отклонение σ, а коэффициент вариации $\gamma = \frac{\mu}{\sigma}$. Набор извлеченных параметров из озвученных фрагментов, содержащих произношение (если не указано иное), выглядит следующим образом:

1. μ, γ высота голоса – логарифмической фундаментальной частоты, F_0, на полутоновой частотной шкале (от 27.5 Гц);

2. μ, γ дрожания – отклонения в длинах индивидуальных последовательных периодов F_0;

3. μ, γ of shimmer – разницы пиковых амплитуд последовательных периодов F_0;

4. μ, γ громкости – воспринимаемой интенсивности сигнала;

5. μ, γ частоты формант 1, 2 и 3 – центральной частоты первой, второй и третьей форманты;

6. μ, γ диапазона форманты 1;

7. μ, γ отношения гармоник к шуму – отношения энергии гармоничных фрагментов к энергии шумоподобных фрагментов;

8. μ, γ альфа отношения – отношения суммарной энергии в промежутке 50–1000 Hz к суммарной энергии в промежутке 1–5 kHz;

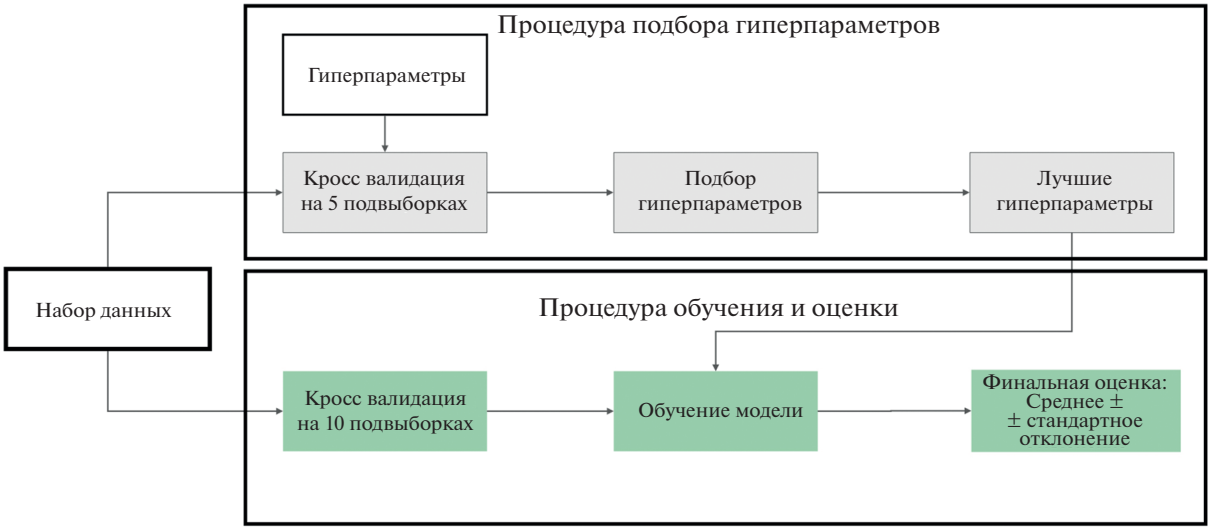

Рис. 1. Процедура вычислений.

9. μ альфа отношения на фрагментах без произношения;

10. μ, γ индекса Хаммарберга – отношение сильнейшего пика энергии в промежутке 0–2 kHz к сильнейшему пику энергии в промежутке 2–5 kHz;

11. μ индекса Хаммарберга на фрагментах без произношения;

12. μ, γ спектрального наклона – линейная регрессия; наклон логарифмического спектра мощности в промежутке 0–500 Hz и 500–1500 Hz;

13. μ спектрального наклона;

14. μ, γ относительной энергии формант 1, 2 и 3;

15. μ, γ гармонической разности H1–H2 – отношения энергии первой F_0 гармоники (H1) к энергии второй F_0 гармоники (H2);

16. μ, γ гармонической разности H1–A3 – отношения энергии первой F_0 гармоники (H1) к энергии наивысшей гармоники в области третьей форманты (A3);

17. частота пиков громкости – количество пиков громкости в секунду;

18. μ, σ суммарной длины фрагментов с произношением;

19. μ, σ суммарной длины фрагментов с без произношения;

20. число продолжительных фрагментов с произношением в секунду;

21. μ, γ мел-кепстральных коэффициентов, MFCC, 1–4 (на всех фрагментах + только на фрагментах с произношением);

22. μ, γ спектрального потока (на всех фрагментах + только на фрагментах с произношением) – разности спектров двух последовательных отрезков;

23. μ, γ диапозона формант 2–3;

24. μ эквивалентного уровня звука.

Процедура вычислений

Процедура вычислений представлена на рис. 1. Для подбора гиперпараметров была использована байесовская оптимизация.

Метрики оценки

Пусть $A = \{a_i\}_{i=1}^N$ – множество целевых переменных, $B = \{b_i\}_{i=1}^N$ – соответствующие предсказанные значения. Мы используем среднюю абсолютную ошибку, $\text{MAE} = \frac{1}{|N|}\sum_{i\in N}|a_i - b_i|$, чтобы оценить, насколько большой, в среднем, может быть ошибка при предсказании. MAE чувствительна к выбросам; чтобы решить эту проблему, мы вычисляли среднюю оценку абсолютной процентной ошибки, MEAPE, следующим образом. Количество итераций $K = 100$, на каждой итерации мы равномерно выбирали M случайных значений из каждого множества A и B и вычисляли среднее значение для каждого: $\bar{A}_k = \frac{1}{M}\sum_{j=1}^M a_i$, $\bar{B}_k$ вычислялось аналогично. Затем мы вычисляли относительные абсолютные ошибки, $\frac{1}{100}\sum_{k=1}^{100}|(\bar{A}_k - \bar{B}_k)/(\bar{A}_k)|$; и фиксировали их среднее и стандартное отклонение. Это позволило количественно оценить среднее и стандартное отклонение качественности подмножества предсказаний.

Основной целью данной работы было проведение исчерпывающего набора экспериментов для эмпирического изучения эффективности различ-

Таблица 1. Области значений гиперпараметров MLP и соответствующие подобранные значения

	N_n	N_e	lr	активация	оптимизатор
	{2,3,...200}	[100, 50 000]	[1e-6, 1e-2]	{Identity, Logistic, Tanh, ReLu}	{LBFGS, SGD, ADAM}
MLP	200	100	–	relu	LBFGS

ных методов регрессии для нахождения надежного решения на основе искусственного интеллекта. Мы изучили эффективность трех семейств моделей: (a) линейные: линейная регрессия; (b) непараметрические: случайный лес, адаптивный бустинг, градиентный бустинг, метод k-ближайших соседей; (c) нейронные сети: многослойный перцептрон, сверточная нейронная сеть, предобученная сверточная нейронная сеть. Поскольку линейная регрессия на отобранных вручную признаках, а также сверточная нейронная сеть и предобученная сверточная нейронная сеть на спектрограммах не дали обнадеживающих результатов, мы не рассматриваем их в данной статье.

Нейронные сети

Пусть $\mathbf{x} \in \mathcal{X}$ и $\mathbf{y} \in \mathcal{Y}$, – объекты и целевые переменные соответственно. Целью является выявить условное распределение вероятностей $p(\mathbf{y}|\mathbf{x},\theta)$ на основе выборки для обучения, $\mathcal{D}=\{(\mathbf{x}_i,\mathbf{y}_i)\}_{i=1}^N$, где N – это размер выборки, а θ обозначает параметры модели, которые предстоит оценить.

Многослойный перцептрон, MLP, настраивает веса $\mathbf{W}_\ell$ и смещения $\mathbf{b}_\ell$ (для $\ell = 1,...,L$) композиции L скрытых слоев для получения распределения функции отображения между входными данными $\mathbf{x}$ и целевыми переменными $\mathbf{y}$, т.е. $p(\mathbf{y}|\mathbf{x};\theta)$, где $\theta = (\mathbf{W}_1,\mathbf{b}_1,...,\mathbf{W}_L,\mathbf{b}_L)$. Конкретно, обозначая скрытые единицы на слое ℓ как $\mathbf{z}_\ell$ и поэлементную (не)линейную функцию активации как $\psi:\mathbb{R}\to\mathbb{R}$, получим:

$$\mathbf{z}_\ell = f_\ell(\mathbf{z}_{\ell-1}) = \psi_\ell(\mathbf{W}_\ell\mathbf{z}_{\ell-1} + \mathbf{b}_\ell). \quad (1)$$

Следовательно, мы можем записать композицию как:

$$f(\mathbf{x};\theta) = f_L(f_{L-1}(....(f_1(\mathbf{x}))...)). \quad (2)$$

где, по договоренности, $\mathbf{z}_1 = \mathbf{x}$.

На каждом слое этой композиции градиенты вычисляются в соответствии с их параметрами по правилу дифференцирования сложной функции, после чего эти градиенты (или производные более высокого порядка) передаются в оптимизатор для настройки параметров. Более подробное описание содержится в [11]. Основные гиперпараметры: (i) число скрытых слоев, (ii) число нейронов N_n, (iii) число эпох N_e, (iv) функции активации, (vi) коэффициент скорости обучения lr, и (vi) оптимизатор. В силу небольшого размера нашего набора данных, чтобы избежать переобучения, мы ограничились неглубокими сетями и использвали только один скрытый слой, а также ограничили размер батча 32. Таблица 1 демонстрирует области значений параметров и соответствующие подобранные значения.

Ансамблевое обучение

Дерево решений (DT) – это иерархическая структура дерева, которая состоит из корневого узла, внутренних узлов и листовых узлов. Корневой узел представляет весь набор данных. Листовые узлы представляют все возможные результаты, полученные этом наборе данных. DT стремится создать наиболее чистые листовые узлы. Для этого DT рекурсивно и жадно ищет комбинацию всех признаков и их значений, чтобы найти лучшую точку разделения, и рекурсия завершается, когда выполняется условие остановки. Подробности см. в [11].

Деревья решений имеют несколько преимуществ, включая легкость интерпретации, быстрое обучение и относительную устойчивость к выбросам. Однако они склонны к переобучению и дают оценку с высокой дисперсией. Предварительный и пост-прунинг, т.е. контроль глубины и ширины дерева, являются популярными техниками для предотвращения переобучения. Однако снижение дисперсии требует более сложных подходов. Один из способов – использовать ансамбль деревьев, например, случайные леса, RF, [8]. RF сначала создает различные выборки бутстрэпа из обучающего набора данных и обучает неподрезанное дерево на каждой выборке, а затем агрегирует прогнозы, усредняя их. Обобщенная модель ансамбля из M деревьев имеет следующий вид:

$$t(y|\mathbf{x}) = \frac{1}{M}\sum_{m\in M}\alpha_m t_m(y|\mathbf{x}), \quad (3)$$

где t_m – m-е дерево, α_m – соответствующий вес. Можно рассматривать это как аддитивную линейную модель с адаптивными базисными функциями. Таким образом, мы можем использовать метод наискорейшего спуска с поиском линии и алгоритмы бустинга. Адаптивный бустинг, AB и градиентный бустинг, GB, основаны на этом принципе. Они последовательно обучают слабую модель, на каждом шаге взвешивают данные, чтобы скорректировать ошибки текущей модели и, наконец, объединяют слабые модели для построения сильной.

Таблица 2. Области значений гиперпараметров AB, RF и GB и соответствующие подобранные значения

	N_e	M_{ss}	M_{sl}	lr	α
AB	{10,11,...,10 000}	–	–	[1e-3, 5e-1]	–
RF	{10,11,...,10 000}	{2,3,...,10}	{1,2,...,10}	–	–
GB	{10,11,...,10 000}	{2,3,...,10}	{1,2,...,10}	[1e-3, 5e-1]	[1e-1, 9e-1]
AB	9533	–	–	–	0.208
RF	5516	2	1	–	–
GB	10 000	5	9	0.087	0.661

Количество моделей N_e, минимальное число образцов, требуемых для разделения внутреннего узла M_{ss}, минимальное число образцов, необходимых в листовом узле M_{sl}, коэффициент скорости обучения lr (для AB и GB) и α альфа-квантиль функции потерь Хьюбера – наиболее важные гиперпараметры. В табл. 2 представлены области значений гиперпараметров и подобранные значения.

Метод k-ближайших соседей

Метод k-ближайших соседей (KNN) [7] предсказывает целевое значение точки данных $\mathbf{x}$ путем определения распределения целевых значений K ее ближайших соседей в обучающем наборе, $N_K(\mathbf{x}, \mathcal{D})$. Конкретно,

$$p(y = c \,|\, \mathbf{x}, \mathcal{D}) = \frac{1}{K} \sum_{n \in N_K(\mathbf{x}, \mathcal{D})} \mathbb{I}(y_n = c), \tag{4}$$

где $\mathbb{I}$ – индикаторная функция.

У KNN два основных гиперпараметра: число ближайших соседей $K \in \{1,2,...,10\}$ и метрика расстояния, определяющая окрестность $\mathbf{x}$. В наших экспериментах мы использовали расстояние Минковского и рассматривали его параметр $P \in \{1,2,...,5\}$ как гиперпараметр. KNN показал лучший результат при $K = P = 1$.

Важность признаков

Мы определили важность признаков для предсказания нашего лучшего оценщика, используя подход Shapley Additive exPlanation (SHAP) и его библиотеку на языке Python [10]. SHAP связывает оптимальное распределение выигрышей с локальными объяснениями, используя значения Шепли из теории кооперативных игр и их связанные расширения. В терминах машинного обучения каждый признак заданного набора данных рассматривается как игрок; игроки могут вести переговоры и формировать коалиции. В случае полного перебора важность каждого признака a для регрессии объекта x вычисляется как среднее по всем возможным комбинациям этого признака с подмножествами S всех остальных признаков относительно выбранной функции значения, как показано ниже [2]:

$$\varphi_a(x) = \sum_{S \subseteq \{1,\ldots,m\} \setminus \{a\}} \frac{|S|!(m - |S| - 1)!}{m!} \times \\ \times (v(x, S \cup \{a\}) - v(x, S)), \tag{5}$$

где m – общее количество признаков, v – выбранная функция значения.

В простейшем случае значение функции v является бинарным, равным 1 для победных коалиций и 0 в противном случае. Если коалиция $S \cup \{a\}$ является победной, в то время как S нет, признак a получает ненулевое значение важности. Однако для больших наборов признаков прямой подход уже не применим из-за комбинаторного взрыва в терминах количества возможных коалиций, и значение функции выражается через приближенное ожидание, вычисленное, например, с помощью метода Монте-Карло.

Мы применили два инструмента SHAP: график-столбцы средних абсолютных значений SHAP (MAS) для каждого признака и график-распределение BeeSwarm Summary (BSS). MAS, в среднем, количественно характеризует вклад каждого признака в предсказанные целевые значения. Чем выше значение MAS для признака, тем выше его влияние. Строки этих двух графиков представляют признаки набора данных, упорядоченные по убыванию сверху вниз. В каждой строке BSS точки распределены горизонтально в соответствии с их значением SHAP; в местах с высокой плотностью значения SHAP стекаются вертикально. Изучение распределения значений SHAP демонстрирует влияние признака на предсказания.

3. РЕЗУЛЬТАТЫ

Наш набор данных состоит из образцов речи, полученных при выполнении трех различных заданий: (i) рассказ по рисункам, (ii) личная история и (iii) инструкция. Чтобы изучить эффективность каждого из заданий, мы применили вышеописанные методы к данным, полученным в каждом из заданий по отдельности, а затем на всех данных без разделения по типу выполненного задания. Результаты регрессии, предсказываю-

Таблица 3. Результаты регрессии: среднее и стандартное отклонение метрик качества на 10 различных разбиениях данных. Лучшие результаты выделены жирным шрифтом

	Все задания		Личная история		Инструкция		Рассказ по рисункам	
	MAE	MEAPE	MAE	MEAPE	MAE	MEAPE	MAE	MEAPE
Случайное предсказание	1.257 ± 0.04	98.637 ± 12.299	1.260 ± 0.073	100.123 ± 20.114	1.26 ± 0.073	100.123 ± 20.114	1.26 ± 0.073	100.123 ± 20.114
К-ближайшие соседи	**0.120 ± 0.056**	**6.790 ± 2.466**	0.703 ± 0.088	27.017 ± 6.605	0.738 ± 0.064	31.208 ± 6.157	0.742 ± 0.071	25.759 ± 3.285
Случайный лес	0.555 ± 0.040	12.111 ± 3.411	**0.697 ± 0.043**	27.176 ± 5.511	**0.673 ± 0.069**	26.530 ± 6.483	**0.713 ± 0.032**	26.419 ± 3.427
Градиентный бустинг	0.524 ± 0.039	12.324 ± 2.926	0.698 ± 0.040	25.151 ± 3.506	0.744 ± 0.060	31.383 ± 7.435	0.779 ± 0.048	29.393 ± 3.392
Адаптивный бустинг	0.654 ± 0.020	14.935 ± 2.808	0.723 ± 0.071	26.977 ± 5.044	0.704 ± 0.054	27.938 ± 4.983	0.736 ± 0.044	**25.183 ± 3.140**
Многослойный перцептрон	0.648 ± 0.069	16.197 ± 4.225	0.718 ± 0.011	**23.394 ± 2.780**	0.712 ± 0.019	**26.448 ± 4.301**	0.732 ± 0.036	33.522 ± 8.632
Mean ± Std.	0.500 ± 0.197	12.453 ± 3.237	0.708 ± 0.011	25.943 ± 1.474	0.714 ± 0.025	28.7014 ± 2.184	0.740 ± 0.022	28.040 ± 3.102

щей тяжесть симптомов депрессии по шкале от 0 (полное отсутствие симптомов) до 3 (наиболее тяжелые симптомы), представлены в табл. 3. Необходимо отметить, что для краткости мы приводим подобранные гиперпараметры только для случая рассмотрения всех данных без разделения по типу задания. Остальные данные доступны по запросу.

При использовании всех данных без деления по типу задания KNN с параметрами $K = P = 1$ демонстрирует лучшие результаты, следующая лучшая модель – градиентный бустинг с параметрами $N_e = 10\,000$, $M_{ss} = 5$ и $M_{sl} = 10$. Хотя все разделения на обучающую и тестовую выборки были непересекающимися, можно ожидать, что некоторые акустические признаки были похожими во всех трех аудио-образцах одного участника. Это объясняет, почему результаты регрессии на всех данных значительно лучше, чем при делении данных по типу задания. Тем не менее результаты, полученные для каждой отдельной задачи, также являются приемлемыми. Лучший MAE демонстрирует метод случайного леса, а MEAPE – многослойный перцептрон. Также важно отметить, что из последней строки этой таблицы можно сделать вывод о наибольшей информативности данных личной истории.

На рис. 2 представлены столбчатый график MAS и BSS.

σ MFCCV4 имеет значение MAS = 0.110, что составляет почти 16% суммарной SHAP-доли, и является наиболее значимым признаком. σ и μ гармонической разницы H1–A3, имеют значения MAS, равные 0.058 и 0.03 и составляют 12.7% суммарной SHAP-доли, что делает гармоническую разницу H1–A3 вторым по важности признаком. σ хрипоты и частоты F1 со значениями MAS, равными 0.033 и 0.028, составляют 4.8% и 4.1$ SHAP-доли и являются третьим и четвертым по важности признаком соответственно.

4. ЗАКЛЮЧЕНИЕ И ПЕРСПЕКТИВЫ

Данное исследование преследовало две цели: (i) разработать надежное, основанное на искусственном интеллекте решение для выявления депрессии у русскоговорящих пациентов на основе аудиозаписей речи; (ii) сопоставить результаты, полученные с помощью объяснимого искусственного интеллекта, с имеющимся знанием в области психиатрии.

Для достижения первой цели мы представили новый уникальный набор данных, состоящий из аудиоданных 247 участников. Мы изучили работу восьми моделей регрессии: сперва мы извлекли спектрограммы аудиофайлов и передали их в сверточные нейронные сети с различными архитектурами. Однако такой подход не принес приемлемых результатов. Развивая это направление,

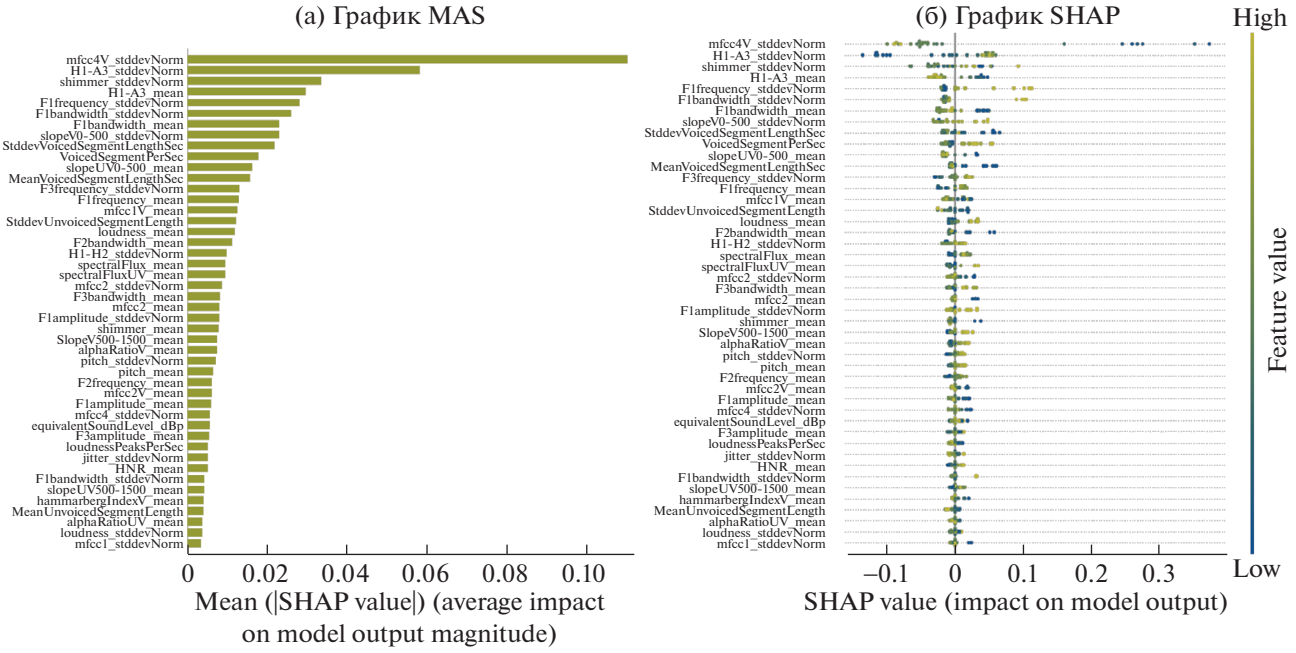

Рис. 2. Строки этих двух графиков представляют объекты, они ранжированы в порядке убывания сверху вниз относительно их значений MAS.

мы использовали различные предобученные нейронные сети, такие как Inception V3, VGG и т.д.; но это тоже не принесло положительных результатов.

По этой причине мы сфокусировались на отобранных вручную признаках и извлекли модифицированный набор eGeMAPS. После этого мы изучили работу пяти методов машинного обучения на четырех вариантах нашего датасета, которые были получены путем деления данных по типу задания: (i) все задания, (ii) рассказ по рисункам, (iii) личная история, (iv) инструкция.

В связи с ожидаемыми сходствами некоторых акустических признаков наш лучший метод регрессии, KNN, достигает приемлемых результатов с MAS = 0.12 и MEAPE = 6.8% на данных всех заданий.

Хотя после разделения аудиофайлов для каждой задачи производительность рассматриваемых методов снизилась, они все равно были приемлемыми. Например, для данных, основанных на личном вопросе, метод случайного леса получил MAE = 0.7, а многослойный перцептрон получил MEAPE = 23.4.

Разделение аудиоданных для каждой задачи не только является надежной основой для оценки качества полученных результатов (так как у каждого участника была только одна строка извлеченных акустических признаков), но и является частью нашего плана по изучению того, какая из трех задач элиситации речи более эффективна. Поскольку, в среднем, все пять рассматриваемых методов показали лучшие результаты на данных, основанных на личном вопросе, мы приходим к выводу, что эта задача более эффективна. Мы также исследовали влияние признаков на производительность метода случайного леса на этих данных с помощью метода SHAP. Наше исследование показало, что MFCCV4, гармоническая разница H1–A1 и хрипота являются тремя наиболее важными признаками.

Настоящая работа имеет ограничения, и они определяют наши будущие направления исследования: (i) Улучшение точности наших прогнозов; (ii) Рассмотрение более продвинутых моделей искусственного интеллекта (таких как трансформеры), в нашем исследовании; (iii) Моделирование аудиоданных без записи реальной речи; (iv) Преобразование аудиофайлов в текст для использования передовых моделей естественного языка и изучения связи между симптомами депрессии и словарным запасом участников.

ВКЛАД АВТОРОВ

С.Ш.: идея, методология, исследование, программирование, валидация, формальный анализ, подготовка черновика статьи, подготовка окончательного текста. К.А.: исследование, программирование, валидация, подготовка черновика статьи. Ш.Т.: курирование данных. Х.М.: идея, курирование данных, подготовка окончательного текста. Д.О.: идея, подготовка окончательного текста, ресурсы, администрирование проекта.

Все авторы прочли и одобрили окончательный текст статьи.

ИСТОЧНИК ФИНАНСИРОВАНИЯ

Публикация подготовлена за счет средств гранта на поддержку исследовательских центров в сфере искусственного интеллекта, в том числе в области "сильного" искусственного интеллекта, систем доверенного искусственного интеллекта и этических аспектов применения искусственного интеллекта, предоставленного АНО "Аналитический центр при Правительстве Российской Федерации" в соответствии с соглашением о предоставлении субсидии (идентификатор соглашения о предоставлении субсидии 000000D730321P5Q0002) и договором с ФГАОУ ВО "Национальный исследовательский университет "Высшая школа экономики" от 2 ноября 2021 г. № 70-2021-00139.

СПИСОК ЛИТЕРАТУРЫ

1. Depressive disorder. https://www.who.int/news-room/fact-sheets/detail/depression
2. *Strumbelj E., Kononenko I.* Explaining prediction models and individual predictions with feature contributions. Knowl. Inf. Syst. 2014. V. 41. № 3. P. 647–665.
3. *Eyben F., Wöllmer M., Schuller B.* opensmile - the munich versatile and fast open-source audio feature extractor. In Proc. ACM Multimedia (MM), ACM, Florence, Italy, 2010. P. 1459–1462.
4. *Eyben F., Scherer K.R., Schuller B.W., Sundberg J., André E., Busso C.,Devillers L. Y., Epps J., Laukka P., Narayanan S.S., et al.* The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. IEEE transactions on affective computing. 2015. V. 7. № 2. P. 190–202.
5. *Mockus J., Tiesis V., Zilinskas A.* The application of Bayesian methods for seeking the extremum. Towards global optimization. 1978. V. 2. № 117–129. P. 2.
6. *Friedman J.H.* Greedy function approxmation: a gradient boosting machine. Annals of statistics. 2001. P. 1189–1232.
7. *Bentley J.L.* Multidimensional binary search trees used for associative searching. Communications of the ACM. 1975. V. 18. № 9. P. 509–517.
8. *Breiman L.* Random forests. Machine learning. 2001. V. 45. № 1. P. 5–32.
9. *Khudyakova M., Antonova N., Nelubina M., Surova A., Vorobyova A., Minnigulova A., Gronskaya N., Yashin K., Medyanik I., Shishkovskaya T., et al.* Discourse diversity database (3d) for clinical linguistics research: Design, development, and analysis. Bakhtiniana. Revista de Estudos do Discurso. 2023. V. 18. № 1. P. 32–57.
10. *Lundberg S.M., Lee S.* A unified approach to interpreting model. predictions. In Advances in Neural Information Processing Systems 30. 2017. P. 4765–4774.
11. *Murphy K.P.* Probabilistic machine learning: an introduction. MIT press, 2022.
12. *Wu P., Wang R., Lin H., Zhang F., Tu J., Sun M.* Automatic depression recognition by intelligent speech signal processing: A systematic survey. CAAI Transactions on Intelligence Technology, 2022.
13. *Hastie T., Rosset S., Zhu J., Zou H.* Multi-class adaboost. Statistics and its Interface. 2009. V. 2. № 3. P. 349–360.

AN EXPLAINED ARTIFICIAL INTELLIGENCE-BASED SOLUTION TO IDENTIFY DEPRESSION SEVERITY SYMPTOMS USING ACOUSTIC FEATURES

S. Shalileh[a,b], A. O. Koptseva[b], T. I. Shishkovskaya[c], M. V. Khudyakova[a,d], and O. V. Dragoy[a,e]

[a]*Center for Language and Brain, HSE University, Moscow, Russia*
[b]*Vision Modeling Laboratory, HSE University, Moscow, Russia*
[c]*Department of Endogenous Mental Disorders and Affective States, Federal State Budgetary Scientific Institution Mental Health Research Center, Moscow, Russia*
[d]*Center for Language and Brain, HSE University, Nizhny Novgorod, Russia*
[e]*Institute of Linguistics, Russian Academy of Sciences, Moscow, Russia*

Presented by Academician of the RAS A.L. Semenov

This paper represents our research to (i) propose an artificial intelligence, AI-based solution to identify depression and (ii) investigate our psychiatric knowledge. Concerning the first objective, we collected and annotated a new audio data set, and scrutinized the performance of eight regression approaches. Our studies showed that k-nearest neighbor and random forest form the group having the most acceptable results. Regarding our second objective, we determined the importance of the features of our best model using the SHapley Additive exPlanations approach: our findings showed that the fourth Mel-frequency cepstral coefficients, harmonic difference, and shimmer are the most important features.

Keywords: depression recognition, acoustic features, regression, explainable artificial intelligence, artificial intelligence