**УДК** 621.391:004.934.2 **DOI** 10.15622/ia.2021.3.1

А.Н. Величко, А.А. Карпов

# АНАЛИТИЧЕСКИЙ ОБЗОР СИСТЕМ АВТОМАТИЧЕСКОГО ОПРЕДЕЛЕНИЯ ДЕПРЕССИИ ПО РЕЧИ

*Величко А.Н., Карпов А.А.* **Аналитический обзор систем автоматического определения депрессии по речи.**

**Аннотация.** В последние годы в медицинской и научно-технической среде возрос интерес к задаче автоматического определения наличия депрессивного состояния у людей. Депрессия является одним из самых распространенных психических заболеваний, непосредственно влияющих на жизнь человека. В данном обзоре представлены и проанализированы работы за последние два года на тему определения депрессивного состояния у людей. Приведены основные понятия, относящиеся к определению депрессии, описаны как одномодальные, так и многомодальные корпусы, содержащие записи информантов с установленным диагнозом депрессии, а также записи контрольных групп, людей без депрессии.

Рассмотрены как теоретические исследования, так и работы, в которых описаны автоматические системы для определения депрессивного состояния — от одномодальных до многомодальных. Часть рассмотренных систем решает задачу регрессивной классификации, предсказывая степень тяжести депрессии (отсутствие, слабая, умеренная, тяжелая), а другая часть – задачу бинарной классификации, предсказывая наличие заболевания у человека или его отсутствие. Представлена оригинальная классификация методов вычисления информативных признаков по трем коммуникативным модальностям (аудио, видео и текстовая информация). Описаны современные методы, используемые для определения депрессии в каждой из модальностей и в совокупности. Наиболее популярными методами моделирования и распознавания депрессии в рассмотренных работах являются нейронные сети. В ходе аналитического обзора выявлено, что основными признаками депрессии считаются психомоторная заторможенность, которая влияет на все коммуникативные модальности, и сильная корреляция с аффективными величинами валентности, активации и доминации, при этом наблюдается обратная корреляция между депрессией и агрессией. Выявленные корреляции подтверждают взаимосвязь аффективных расстройств с эмоциональными состояниями человека. В множестве рассмотренных работ наблюдается тенденция объединения модальностей для улучшения качества определения депрессии.

**Ключевые слова:** автоматическое определение депрессии, компьютерная паралингвистика, речевые технологии, машинное обучение

**1. Введение.** Согласно данным ВОЗ [1], депрессия является распространенным психическим расстройством и одной из основных болезней, которые приводят к ухудшению жизнедеятельности человека, и может стать причиной инвалидности. На 2018 год во всем мире около 264 млн. человек во всех возрастных группах страдали от депрессии [2].

В последние 10 лет возрос интерес к системам автоматического определения депрессии. На это повлияли многие причины — тяжесть заболевания и повсеместная распространенность, отсутствие лабораторных тестов или процедур для диагностики депрессии и так далее. На данный

момент наличие заболевания определяется путем беседы со специалистом-психотерапевтом и заполнения различного рода опросников: состояния здоровья (Patient Health Questionnaire, PHQ) [3], шкала депрессии Бека (Beck Depression Inventory) [4], самооценки депрессивных симптомов (Quick Inventory of Depressive Symptoms - Self Report, QIDS-SR) [5], шкала Гамильтона для оценки депрессии (Hamilton Rating Scale for Depression, HRSD [6] и других. Кроме того, профессиональная оценка может ва-рьироваться в зависимости от компетентности специалиста и методов диагностики, которые он использует. Таким образом, на данный момент не существует объективного метода диагностики депрессии.

Многие работы представляют автоматические системы для определения состояния депрессии — существуют как одномодальные, так и многомодальные системы. Кроме того, часть систем решает задачу регрессии (определяя степень тяжести заболевания), а часть — задачу бинарной классификации (для определения наличия заболевания или его отсутствия). Задача определения депрессии была неоднократно представлена на соревнованиях AVEC (Audio-Visual Emotion Challenge) в 2013 [7], 2014 [8], 2016 [9], 2017 [10] и 2019 годах [11].

Отмеченные в данной работе системы опираются на гипотезу о том, что эмоциональное состояние диктора существенно влияет на акустические характеристики (спектральные и просодические) его речи. Лингвистические и нелингвистические факторы влияют на фонетические характеристики речи. Среди таких факторов можно отметить: физическое и психическое состояния говорящего, различные патологии мышления и психические болезни, ряд болезней, влияющих непосредственно на возможность речеобразования, и другие [12].

Данная работа проведена с целью показать значимость существующих проблем, которые вызваны депрессией, описать подходы, которые используются специалистами, и выполнить аналитический обзор существующих автоматических систем определения депрессии.

В разделе 2 приводится определение депрессии, влияние заболевания на различные аспекты жизни человека, описание используемых на данный момент специалистами методов выявления депрессии. В разделе 3 описаны существующие базы данных, на основе которых возможно обучение автоматической системы определения депрессии. Одно- и многомодальные системы, представленные за последние два года, описаны в разделах 4 и 5. Раздел 6 содержит выводы, сделанные в результате аналитического обзора, также представлена классификация наиболее эффективных методов для построения таких систем. На основе проведенного анализа в разделе 6 также описаны требования, которые могут

быть выдвинуты к автоматическим системам определения депрессии. Выводы по проделанной работе описываются в разделе 7.

**2. Основные определения и медицинские характеристики депрессии.** Построение автоматической системы включает в себя в том числе и понимание рассматриваемой задачи. В случае с определением депрессии возможна векторизация и использование некоторых признаков депрессии для обучения моделей, а именно тех признаков, которые проявляются вербально и невербально и на которые обращает внимание специалист при личной беседе с пациентом.

Аффективные состояния — это психические состояния, которые характеризуются заметной эмоциональной окрашенностью: эмоциональные состояния, состояние аффекта, настроение и тому подобное. Изменения аффективного состояния являются естественной характеристикой поведения людей. Однако, когда эти изменения становятся интенсивными, длятся продолжительное время и при этом ухудшается жизнедеятельность человека, есть вероятность, что может проявиться аффективное расстройство. В отличие от кратковременных эмоций, настроение — длительное по времени аффективное состояние и, следовательно, клиническая депрессия – это расстройство настроения, которое может длиться неделями, месяцами и даже годами, изменяясь в тяжести заболевания, если не получено соответствующее лечение.

Расстройства настроения, несомненно, касаются естественных эмоциональных состояний. В частности, схема поведения людей, страдающих от таких расстройств настроения, как униполярная депрессия, показывает сильную временную корреляцию с аффективными величинами валентности, активации и доминации [8]. Данные аффективные величины используются для определения эмоционального состояния человека.

Специалисты выделяют два противоположных аффективных расстройства (расстройств настроения): депрессию (или большое депрессивное расстройство, БДР) и манию [1]. В психиатрии они обозначаются при помощи терминов униполярная депрессия (пациентов беспокоит депрессивное состояние) и биполярное аффективное расстройство (БАР, пациенты переживают как депрессию, так и манию). При этом депрессия и мания могут проявляться одновременно, что приводит к смешанному аффективному эпизоду. Кроме того, мания и депрессия могут проявляться в менее тяжелой форме (гипомания и дистимия соответственно) или могут быстро сменяться, что называют быстрой циркуляцией фаз. На рисунке 1 показаны течения описанных заболеваний, приведенные в работе [13] (для удобства восприятия два графика течения заболеваний были объединены в один, переведены надписи на графиках, а также график был представлен

в монохромном виде). Верхняя граница рисунка обозначает состояние мании, нижняя указывает на состояние депрессии, а средняя – нормальное состояние. Пунктиром показаны менее тяжелые состояния гипомании и дистимии, которые также являются отклонением от нормального состояния, по горизонтальной оси указано течение времени, а по вертикальной – валентность заболевания.

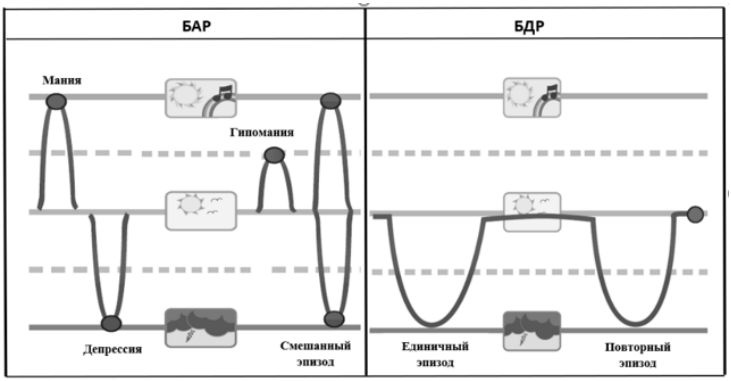


Рис. 1. Фазы течения заболевания при БАР и при БДР [13]

Согласно диагностическому и статистическому руководству по психическим расстройствам V издания (Diagnostic and Statistical Manual of mental disorders V, DSM-V) [14] для постановки диагноза депрессии необходимо, чтобы на протяжении как минимум двух недель присутствовали 5 или более симптомов (включая как минимум один из основных: подавленное настроение и/или потеря интереса и утрата способности получать удовольствие от приятной ранее деятельности):

– подавленное настроение;

– потеря интереса и утрата способности получать удовольствие от приятной ранее деятельности;

– расстройства сна и аппетита;

– психомоторное возбуждение/заторможенность;

– повышенная утомляемость и снижение энергичности;

– сниженная самооценка, чувство никчемности или неадекватное чувство вины;

– снижение способности к концентрации внимания или заторможенное мышление;

– суицидальные тенденции.

Крайней формой выражения депрессии является самоубийство, а риск самоубийства у пациентов с депрессией в течение жизни со-

ставляет 15%. В работе [15] был проведен сравнительный метаанализ, который показал 3428 факторов риска суицида среди 365 лонгитюдных (длительных) исследований. Авторы сделали вывод, что все обозначенные факторы риска недостаточно точно могут предсказать суицид, что частично может быть вызвано методологическими ограничениями существующей литературы. По мнению авторов работы [15], исследования показателей оценивания рисков суицида с использованием клинического инструментария, предсказывающего суицид, также недоработаны, а последние систематические обзоры показывают, что на данный момент нет инструментария для определения риска суицида, который показывал бы высокую точность. В работе [16] проведен статистический анализ на основе англоязычных рецензированных статей. В рассмотренных авторами исследованиях предикативные модели должны были предсказать смерть путем самоубийства или попытку самоубийства. Была проведена серия симуляций на гипотетической популяции индивидов с целью выявления преимуществ в статистическом моделировании, которые могли бы повысить способность предсказания попыток суицида и смертей. Было отмечено, что системы предсказания суицидального поведения должны проходить клинические тесты до внедрения в систему здравоохранения, чтобы показывать высокую точность именно в клинических случаях, а также предполагать более общие результаты, включая суицидальные тенденции, поскольку ошибочные гипотезы системы (как ложноположительные, так и ложноотрицательные) могут навредить пациентам.

Задача определения депрессии относится к задаче определения деструктивных паралингвистических явлений, которые также включают в себя такие явления, как агрессия, ложь и прочее. Связь между агрессией и депрессией в известных теориях выражена неявно. Редкие случаи депрессии, предположительно, были обнаружены в обществах и группах, где агрессия могла быть сразу выражена, например в военных сражениях и там, где процент случаев насильственной смерти велик. Тем не менее немногие межкультурные работы, существующие на момент написания статьи [17], не поддерживали гипотезу об обратной корреляции между агрессией и депрессией. Большинство рассмотренных автором статьи [17] работ подтверждали, что обязательным компонентом депрессии является чувство вины. Несмотря на то, что автор утверждает, что рассмотренные им работы имели методологические недоработки, сочетание признаков депрессии (депрессивное настроение, суточные циркадные изменения, усталость, бессонница, потеря интереса, потеря веса, периодичность и двухфазовая природа заболевания) все-таки обнаруживается во всех культурах.

Перед автоматическими системами определения депрессии, в том числе, стоит задача проверки данной гипотезы, а также использование выявленных корреляций с другими деструктивными паралингвистическими явлениями и аффективными величинами для получения более точного предсказания автоматической системы о наличии или отсутствии заболевания.

**3. Многомодальные базы данных для исследования депрессии.** Для создания автоматической системы определения депрессии необходимо иметь данные для обучения моделей, которые содержат как речь информантов с установленной депрессией, так и информантов, у которых не было депрессии на момент записи. Среди работ, которые будут рассматриваться в данном аналитическом обзоре, встречаются одномодальные и многомодальные корпусы. В таблице 1 представлены систематизированные данные об описанных корпусах, а именно: язык, количество дикторов/авторов текстов, методы оценивания заболевания, количество данных, доступность. Большинство корпусов является многомодальным, однако описаны также и одномодальные корпусы: Mundt и SH2-FS (Free Speech) (аудио), RusNeuroPsych и Текстовый корпус эссе (текст), корпус информации из профилей социальной сети «Вконтакте» (текст, изображения). Кроме того, некоторые многомодальные корпусы содержат лингвистическую информацию: Pitt (ручная транскрипция), DAIC (автоматическая транскрипция), General Psychotherapy Corpus (транскрипция). В корпусе DAIC также содержатся записи сенсора глубины и физиологические данные (гальваническая проводимость кожи, ЭКГ, дыхание). В столбце Язык сокращения Австрал. от Австралийский и Амер. от Американский.

AViD-Corpus [7] содержит 340 аудио- и видеозаписей, где участники (всего 292 участника) взаимодействуют с компьютером, в то время как их действия записываются камерой и микрофоном. Количество записей каждого участника варьируется от 1 до 4. Длина видео составляет от 20 до 50 минут, в среднем — 25 минут. Общая длительность видео составляет 240 часов. Возраст участников — от 18 до 63 лет, средний возраст — 31,5 лет. Поведение участников в ходе записей было определено заданием: произнесение букв, произнесение букв громким голосом, произнесение букв с улыбкой, повышение громкости голоса при выполнении задания.

Участники читали отрывки из новелл и преданий, пели, рассказывали истории из своего прошлого, рассказывали вымышленные истории для тематического апперцепционного теста (апперцепция — восприятие, узнавание на основе прежних переживаний). Аудиоданные были записаны с использованием микрофонной гарнитуры с частотой 41 кГц, 16 бит.

Таблица 1. Сравнение данных в корпусах, содержащих данные людей с депрессией

| Корпус | Язык | Количество человек | Оценка заболеваний | Количество данных | Доступность |
|---|---|---|---|---|---|
| AViD-Corpus [7] | Немецкий | 292 | Шкала Бека-2 | 240 часов | По запросу |
| BlackDog [19] | Австрал. английский регионалект | 30 | QIDS-SR | 509 минут | Не доступен |
| Pitt [20] | Амер. английский регионалект | 19 | HRSD, QIDS-SR | 355 минут | Не доступен |
| Distress Analysis Interview Corpus (DAIC) [21] | Амер. английский регионалект | Не известно | 5 опросников | 621 сессия | По запросу |
| Mundt [23] | Амер. английский регионалект | 35 | Текущее лечение, шкала Гамильтона | Не известно | Не доступен |
| General Psychotherapy Corpus [24] | Амер. английский регионалект | Не известно | Оценка специалиста | 1300 сессий | Платный доступ (30 дней бесплатного пользования) |
| SH2-FS (Free Speech) [25] | Амер. английский регионалект | 887 | PHQ-9 | 16 часов, 5937 аудио файлов | Не доступен |
| RusNeuroPsych [22] | Русский | 447 (246 детей до 18 лет и 209 взрослых) | 3 опросника | 643 (252 текста детей и 392 текста взрослых) | Открыт |
| Текстовый корпус эссе [46] | Русский | 164 | 10 опросников | 164 текста | Не доступен |
| Корпус информации из профилей социальной сети «Вконтакте» [47] | Русский | 1330 | Шкала депрессии Бека | Не известно | Не доступен |

Видеоданные были записаны с использованием различных кодеков и кадровых частот, а затем сконвертированы в формат до 30 кадров в секунду с разрешением 640x480 пикселей, 24 бита на пиксель. Аннотация проводи-лась согласно опроснику депрессии, шкале Бека-2 [18]. Данный опросник является вторым пересмотром опросника Бека, принятым в 1996 году. Он содержит 21 вопрос, где каждый вопрос оценивается по шкале от 0 до 3 в соответствии с тяжестью симптома. Суммарный балл составляет от 0 до 63, при том, чем выше значение,

тем серьезнее симптомы депрессии. Средний уровень депрессии по шкале Бека у информантов данного кор-пуса был 15,1 для обучающего набора и 14,8 — для отладочного набора, стандартное отклонение 12,3 и 11,8 соответственно. Каждой записи было присвоено единственное значение метки наличия депрессии.

BlackDog [19] – англоязычный корпус (австралийский английский регионолект), собранный в Австралии при помощи 1 камеры и 1 микрофона. Частота дискретизации аудио 44,1 кГц. Метки бинарной классификации были проставлены вручную и содержат категории «тяжелая депрессия» и «здоровые субъекты». Всего было записано 30 информантов для каждого класса, количество мужчин и женщин сбалансированное (30 и 30 соответственно). Интервью содержало вопросы с открытым ответом, а симптомы депрессии измерялись при помощи QIDS-SR. Средний показатель тяжести депрессии у информантов данного корпуса — 19 при разбросе от 14 до 26. Общая длительность записей — 509 минут, из которых 119,3 минуты – чистая речь информантов.

Pitt [20] — англоязычный корпус (американский английский регионолект), собранный в Питтсбурге при помощи 4 камер и 2 микрофонов. Частота дискретизации аудио 48 кГц. Метки бинарной классификации содержат категории «тяжелая депрессия» и «легкая депрессия». Кроме того, проводилась ручная транскрипция. Всего было записано 19 информантов для каждого класса, количество мужчин и женщин — 14 и 24 соответственно. Опрос содержал вопросы из клинического интервью HRSD. Средние показатели тяжести депрессии у информантов данного корпуса — 22,4 при диапазоне от 17 до 35 для тяжелой депрессии, и — 2,9 при разбросе от 1 до 7 для легкой депрессии. Эквивалентные показатели по шкале QIDS-SR: средний показатель — 17 при диапазоне от 13 до 26 для тяжелой депрессии, и — 2 при диапазоне от 1 до 5 для легкой депрессии. Общая длительность записей составляет 355 минут, из которых 92 минуты – чистая речь информантов.

Distress Analysis Interview Corpus (DAIC) [21] — многомодальная коллекция клинических интервью. Корпус разработан для симуляции стандартных процессов определения того, имеется ли у человека риск ПТСР (посттравматического стрессового расстройства) и большого депрессивного расстройства. Корпус содержит следующие типы интервью:

– очные интервью (лицом к лицу) между участниками и интервьюером;

– телеконференции, где интервью проводилось с использованием телеконференций;

– «Волшебник Оз» (WoZ) или «Гудвин» (Wizard of Oz, WoZ) – интервью проводилось анимированным виртуальным интервьюером по имени Элли, которую контролировал интервьюер в другой комнате;

– автоматические интервью – интервью проводилось в автоматическом режиме с Элли.

Разработчиками были выбраны две группы жителей Лос-Анджелеса – ветераны военных сил США и гражданские. Все они были проверены опросниками на депрессию, ПТСР и тревожность. Интервью начинались с нейтральных вопросов, затем вопросы становились более специфичными (о симптомах, событиях), а заканчивались фазой спокойствия. Корпус включает в себя аудио-, видеоданные и записи сенсора глубины всех взаимодействий. В коллекции также присутствуют физиологические данные (гальваническая проводимость кожи, ЭКГ, дыхание). Перед и после интервью участники заполняли ряд опросников, включающие базовые вопросы о биографии, измерение психологического стресса, а также измерение текущего настроения. Использовались следующие опросники: The Positive and Negative Affect Scale (PANAS) для оценки настроения, PTSD Checklist – Civilian Version для оценки ПТСР, Patient Health Questionnaire для оценки психического здоровья, Depression module для оценки наличия и уровня депрессии, State-Trait Anxiety Inventory для оценки тревожности.

При экспериментах по обучению системы определения депрессии и ПТСР по записям речи из данного корпуса авторам корпуса удалось добиться точности распознавания депрессии в 75,0%, а ПТСР – 72,0%. В записях с волшебником Оз в данных также были найдены некоторые признаки стресса: участники, испытывающие стресс, медленнее начинали говорить и использовали меньше заполненных пауз, чем участники, не испытывающие стресс. Кроме того, от типа стресса зависело, какие признаки наиболее предсказуемы. Если стандартное отклонение перед началом речи каждого участника диалога изменялось, это было лучшим признаком для предсказания депрессии. Однако для ПТСР более информативным было определение среднего количества заполненных пауз в сегменте. Время перед началом ответа при личных вопросах и длительность речи при ответах на вопросы для установления контакта указывают на наличие стресса. Участники использовали меньше заполненных пауз при диалоге с агентом, чем при диалоге с человеком. Участники выражали меньше страха или негативных проявлений, когда агент представлялся автоматическим, чем когда агент был представлен как управляемый человеком. Кроме того, участники показывали больше эмоций из категории «грусть», когда были уверены, что взаимодействуют с компьютером, а не с человеком.

RusNeuroPsych [22] – текстовый русскоязычный корпус, который содержит 643 текста на русском языке 447 авторов в возрасте от 12 до 35 лет. Корпус разделен на две части – подкорпус «дети» (тексты написаны детьми школьного возраста от 12 до 17) и подкорпус «взрослые» (тексты написаны взрослыми от 18 до 35 лет). Первый подкорпус содержит 252 текста 246 человек, а в метаданных указан их пол, год рождения, исследования их моторного, сенсорного и латерального профиля, а также результаты психологического тестирования, которое использовалось для измерения уровня агрессии, тревожности, ригидности и фрустрации. Второй подкорпус состоит из 392 текстов 209 человек, а в метаданных указан идентификатор участника, пол, год рождения, образование, результаты теста большой пятерки, исследования их моторного, сенсорного и латерального профиля, а также результаты шкалы тревоги и депрессии.

Mundt [23] – корпус содержит записи речи за период в 6 недель 35 информантов (20 женщин и 15 мужчин), чей средний возраст 41,8 лет. Одним из условий записи было начало фармакотерапевтического и/или психотерапевтической лечения депрессии. Участники читали заранее подготовленный текст и описывали свои эмоциональные и физические ощущения. Для оценки наличия или отсутствия депрессии использовалась шкала Гамильтона.

General Psychotherapy Corpus [24] – корпус состоит из 1300 транскрибированных терапевтических сессий, которые покрывают различные клинические подходы. Метаданные представлены на уровне сессий и включают демографические сведения для терапевта и пациента, симптомы, которые испытывает пациент, и сводную информацию, прилагающуюся к каждой сессии, обозначаемую как «название». Каждая сессия состоит из следующих друг за другом коммуникативных ходов, аннотированных как сторона терапевта и сторона клиента. Среди 1262 сессий 881 аннотированы как «без депрессии» и 381 как «с депрессией».

SH2-FS (Free Speech) [25] – содержит записи речи в естественных условиях (дома, на работе, в машине) и оценки по самодиагностическому тесту PHQ-9. Общая длительность аудио в корпусе 16 часов речи 887 участников (436 женщин и 450 мужчин), всего 5937 аудио файлов. На всех записях присутствует фоновый шум.

Текстовый корпус эссе [46] – корпус эссе длиной в одну страницу на тему «Я, другие, мир», записанный с целью определить лингвистические характеристики текстов людей с установленной депрессией и отсутствием депрессии. Также участников просили заполнить 10 опросников. В исследовании участвовали две группы испытуемых: 22 пациента Федерального государственного бюджетного учреждения "Научный центр психического

здоровья"(ФГБНУ НЦПЗ) с установленной депрессией и 142 студента гуманитарных и технических вузов Москвы и Кургана, а также взрослые жители этих городов. Корпус информации из профилей социальной сети «Вконтакте» - данные профилей были собраны в период с января 2017 по апрель 2019, также были получены баллы участников по шкале депрессии Бека. Всего была получена информация 1330 профилей, 425 мужчин и 904 женщин, средний возраст 25 лет, а средний показатель по шкале Бека 18,79. Также в корпусе имеется 485121 изображений, собранных из альбомов, аватаров и постов в профилях социальной сети «Вконтакте» 398 волонтеров. В контрольной группе был 201 профиль, а в группе с депрессией – 197.

**4. Обзор работ, представленных на соревнованиях AVEC.** В рассмотренных далее работах в основном исследуется униполярная депрессия или БДР. Значительная часть работ представлена в рамках соревнований по аудиовизуальному определению эмоций AVEC. В общем виде процесс построения автоматической системы для многомодального определения паралингвистических явлений заключается в следующем: имеется многомодальная база данных, содержащая видеозаписи информантов с диагностированной депрессией, и информантов с отсутствием депрессии, а также лингвистическая составляющая их речи. Из данных каждой модальности вычисляются информативные признаки, которые впоследствии являются входными данными для классификатора/регрессора. Выходными же данными классификатора/регрессора являются либо метка класса, либо единственное значение регрессора для классификации и для регрессивной классификации соответственно. Лучшие результаты систем, представленных на соревнованиях 2019 года, отображены в таблице 2. В ней указаны авторы системы, модальности и признаки, которые были использованы для обучения моделей, а также сами модели и результаты классификации, которых удалось добиться авторам по показателям точности CCC (Concordance Correlation Coefficient [27]) и RMSE (Root Mean Squared Error [27]). В качестве сокращений указаны Dev – набор данных для разработки (Development set), а Test – набор данных для тестирования (Test set).

В соревнованиях 2019 года [29] были представлены следующие темы: определение настроения (State-of-Mind Sub-challenge), использование искусственного интеллекта для определения депрессии (Detecting Depression with AI Sub-challenge) и определение эмоций в разных культурах (Cross-cultural Emotion Sub-challenge). В качестве набора данных для определения наличия депрессии был представлен корпус E-DAIC, расширенная версия WOZ-DAIC [21].

Таблица 2. Сравнение результатов систем, представленных на соревнованиях AVEC'19

| Работа | Модальность | Признаки/ Классификатор | CCC (Dev / Test) | RMSE (Dev / Test) |
|---|---|---|---|---|
| Базовая работа Ringeval F. et al. [29] | Аудио | MFCCs | 0,198 / - | 7,28 / - |
| | | eGeMAPS | 0,076 / - | 7,78 / - |
| | | BoAW-M | 0,102 / - | 6,32 / - |
| | | BoAW-e | 0,272 / 0,045 | 6,43 / 8,19 |
| | | DS-DNet | 0,165 / - | 8,09/ - |
| | | DS-VGG | 0,305 / 0,108 | 8,00 / 9,33 |
| | Видео | FAUs | 0,115 / 0,019 | 7,02 / 10,0 |
| | | BoVW | 0,107 / - | 5,99 / - |
| | | ResNet | 0,269 / 0,120 | 7,72 / 8,01 |
| | | VGG | 0,108 / - | 7,69 / - |
| | Аудио + Видео | Все признаки | 0,336 / 0,111 | 5,03 / 6,37 |
| Kaya H. et al. [30] | Аудио + Видео | Объединение разработанных классификаторов | 0,481 / 0,344 | - |
| Ray A. et al. [31] | Аудио | Funct MFCC | - / - | 5,11 / - |
| | Видео | BoVW | - / - | 5,70 / - |
| | Текст | Текст | - / - | 4,37 / - |
| Makiuchi M.R. et al. [32] | Аудио | CNN | 0,338 / 0,199 | 5,97 / 7,02 |
| | | GCNN-LSTM | 0,497 / - | 5,70 / - |
| | Текст | LSTM | 0,360 / 0,048 | 4,97 / 6,88 |
| | | 8 CNN blocks-LSTM | 0,685 / - | 4,22 / - |
| | Видео | GCNN | 0,372 / - | 5,74 / - |
| | Аудио + Текст | CNN and LSTM | 0,452 / 0,213 | 5,08 / 6,42 |
| | Аудио + Текст | GCNN-LSTM и 7 CNN blocks-LSTM | 0,696 / 0,403 | 3,86 / 6,11 |
| | Аудио + Текст + Видео | GCNN-LSTM, 7 CNN blocks-LSTM и GCNN | 0,624 / - | 4,86 / - |
| Fan W. et al. [33] | Аудио + Видео + Текст | Ансамбль классификаторов | 0,466 / 0,430 | 5,07 / 5,91 |
| Yin S. et al. [34] | Аудио + Видео + Текст | Иерархическая двунаправленная LSTM | 0,402 / 0,442 | 4,94 / 5,50 |

Аудиопризнаки состояли из следующих наборов: MFCC (Mel-frequency Cepstral Coefficients), eGeMAPS, BoAW (Bag of Audio Words), Deep Spectrum. В качестве видеопризнаков использовались FAU (Face Action Units), BoVW (Bag of Video Words), ResNet признаки. Базовая модель организаторов представляла собой однослойную 64-d (размерную) сеть с управляемыми рекуррентными блоками (Gated Recurrent Units, GRU)

в качестве рекуррентной сети, за которой следует 64-d полносвязный слой для получения единственного значения оценки регрессии. Лучший результат на отладочном наборе данных для аудиомодальности был получен при использовании признаков Deep Spectrum, вычисленных с использованием сети VGG-16, по показа-телю CCC = 0,289. Для видеомодальности лучший результат был получен при использовании ResNet признаков: CCC = 0,269, RMSE = 7,72. При объединении всех репрезентаций результат, полученный на отладочном наборе, был улучшен до значений CCC = 0,336 и 0,111 на отладочном и на тестовом наборах соответственно. Значения RMSE также были улучшены до 5,03 и 6,37 на отладочном и тестовом наборах.

Работа [30] подготовлена нашим коллективом авторов совместно с зарубежными коллегами. В ней мы аугментировали акустические признаки, предложенные организаторами при помощи транскрипций, полученных при помощи автоматического распознавания речи (Automatic Speech Recognition, ASR). Затем эти транскрипции были использованы для простого представления мешка слов (Bag of Words, BoW), после чего применялся анализ главных компонент для регрессии. Авторы использовали продолжительность реплик из транскрипций для получения общей длительности тишины и дыхания для каждого участника. Для моделирования автоматической сегментации на 7 классов (речь виртуального агента, дыхание, эксплетивные слова, звук губ, смех, тишина, речь субъекта) авторы экспериментировали с признаками eGeMAPS и Deep Spectrum (VGG-16), где применяли оконный метод (100, 200, 400, 500, 1000 мс) с наложением в 1 секунду. Супрасегментные признаки были смоделированы при использовании KELM (Kernel Extreme Learning Machine). Наилучший результат (CCC = 0,344) на тестовом наборе был получен при использовании простых признаков на основе транскрипции ASR (подсчет слов, длительность и BoW). Помимо KELM использовалась ELM с взвешенным ядром (Weighted Kernel ELM), которая присваивает веса высокой важности классу с наименьшим количеством объектов обучения и затем пытается максимизировать невзвешенную среднюю полноту (Unweighted Average Recall, UAR). Лучший результат сегментации на отладочном наборе в 4 из 17 аудиофайлов, полученных с 500 мс окном без наложения с использованием функционалов eGeMAPS LLDs и Weighted KELM, был UAR=65,75%. Результаты показали, что, хотя паттерны в части невербальных признаков сигнала важны, объединение их с лингвистической информацией позволяет добиться лучших результатов без использования современных акустических и видеопризнаков.

Авторы работы [31] объединили признаки трех модальностей и использовали многоуровневую сеть с вниманием, которая обучалась взаимосвязям как между модальностями, так и внутри модальностей. Сеть использует несколько низкоуровневых и среднеуровневых признаков из аудио- и видеомодальностей, а также векторные представления предложений (sentence embeddings). В архитектуре предложенной сети контекстные признаки каждой модальности проходят через двуслойные сети прямого распространения, а выходные данные этих трех сетей объединяются в другую последовательную двунаправленную LSTM (Long-Short Term Memory), а именно stacked BLSTM (Bidirectional LSTM). При помощи предложенной системы результат базовой системы был улучшен на 17,52% по показателю RMSE. Отдельные сети с вниманием для аудио- и видеомодальностей превзошли результат базовой системы по показателю UAR на 20,5%, а текстовая модель с вниманием превзошла лучший результат на 8,95%.

В работе [32] авторы использовали аудиомодальность и ее лексическую составляющую для определения депрессии. Для векторизации аудиомодальности использовались признаки deep spectrum из предобученной сети VGG-16 и применялась Gated Convolutional Neural Network (GCNN), а затем LSTM слой. Для получения лингвистических представлений были извлечены признаки BERT и применена CNN, за которой следовал LSTM слой. Извлечение лингвистических признаков проводилось при помощи BERT модели, в которой 24 слоя (блоков трансформера), 1024 скрытых слоя и 16 блоков «self-attention». Таким образом, всего получается 340 млн. параметров, которые извлекаются из последнего слоя BERT и преобразуются в единый массив признаков для каждого токена (word token). Аудиомодель на основе GCNN-LSTM состоит из последовательных управляемых сверточных блоков, после которых идет слой LSTM и полносвязный слой. Лингвистическая модель была построена на основе CNN-LSTM. При использовании предложенного подхода удалось получить результаты CCC = 0,696 на отладочном наборе и CCC = 0,403 на тестовом наборе.

Авторы работы [34] в качестве видеопризнаков использовали: направление взгляда, 3D положение и ориентацию головы, 17 FAUs для каждого фрейма с весами точности. Для аудиомодальности использовались две категории признаков – 4096-размерный вектор, вычисленный из активации второго полносвязного слоя VGG-16, и 1920-размерный вектор, вычисленный из активации последнего слоя со сжатием в DenseNet-201. Для текстовой модальности использовалась RNN с кодировщиком-декодировщиком для создания вектора семантической репрезентации. Для

создания вектора эмоциональной репрезентации использовался словарь NRC Emotion Intensity Lexicon. Предложенный подход включает в себя две иерархические двунаправленные LSTM для объединения многомодальных признаков и предсказания тяжести депрессии. На отладочном и тестовом наборах были достигнуты результаты 0,402 и 0,442 по показателю CCC; 4,94 и 5,50 по показателю RMSE соответственно.

**5. Обзор работ по автоматическому определению депрессии, представленных вне соревнований AVEC.** Результаты систем, представленных вне соревнований AVEC, отображены в таблице 3. В ней указаны авторы системы, модальности, которые были использованы для обучения моделей, а также сами модели. Перечислены различные показатели точности, которые были использованы авторами, и результаты, полученные по этим показателям с использованием разработанных моделей. В работе [43] в столбце Классификатор под Первая модель подразумевается модель SVM, модели tf-idf со стилистическими и морфологическими признаками, а под Вторая модель подразумевается модель SVM, векторные представления слов, стилистические признаки.

В работе [35] для обучения моделей авторы использовали DAIC-WOZ корпус, а именно трехмерные изображения лица и речь информантов. Методика включает в себя использование векторных представлений на уровне предложений (sentence-level “summary"embedding), LSTM и casual-CNN. Предсказывался показатель PHQ и бинарная классификация, о наличии у пациента БДР, обычно со значением PHQ более или равным 10, либо же отсутствие заболевания. Особенность данного метода заключается в том, что предложенная система не полагается на контекст интервью. Кроме того, она принимает на вход сырые данные (аудио, трехмерные модели лиц и транскрипция), которые суммируются в один вектор. Стоит отметить, что в подходе используются сделанные вручную и предобученные векторные представления на уровне слов на входе, то же самое и на уровне предложений. Модель показывает среднюю ошибку MAE = 3,67 по опроснику здоровья (PHQ), а также 83,3% чувствительности (Sensitivity [28]), 82,6% специфичности (Specificity) и F1-меру 76,9%.

В работе [36] также использовался DAIC-WOZ корпус. Авторы предложили следующую архитектуру (CombAtt): кодировщики модальностей, которые принимают унимодальные признаки на вход и выдают закодированные данные; сети с механизмом внимания для объединения сетей отдельных модальностей. Также они предложили сети для регрессии, которые принимают на вход объединенные данные из второго компонента и на выходе предсказывают баллы по шкале PHQ-8. Использование предложенного подхода позволило авторам получить результаты RMSE = 4,14 и MAE = 3,07, а также EVS (explained variance score) = 0,62.

Таблица 3. Лучшие результаты систем, представленных вне соревнований AVEC

| Работа | Модальность | Классификатор | Показатель | Результат |
|---|---|---|---|---|
| Haque A. et al. [35] | Аудио + Видео + Текст | Casual CNN | F1-мера, | 76,9%, |
| | | | Precision, | 71,4%, |
| | | | Recall, | 83,3%, |
| | | | Average Error | 3,67 |
| Qureshi S.A. et al. [36] | Аудио + Видео + Текст | CombAtt network | RMSE, | 4,14, |
| | | | MAE, | 3,34, |
| | | | EVS | 0,62 |
| Niu M. et al. [37] | 1.7cmАудио | Гибридная сеть (CNN, LSTM и DNN) и lp-нормированное сжатие | RMSE, | 9,66, |
| | | | MAE | 8,02 |
| Rohanian M. et al.[38] | Аудио+Видео +Текст | LSTM с механизмом окна | F1-мера, | 81,0%, |
| | | | Precision, | 80,0%, |
| | | | RMSE | 3,61, |
| | | | MAE, | 4,99 |
| Tao F. et al. [39] | Аудио | SVM | Accuracy | 84,5% |
| Xezonaki D. et al. [40] | Текст | Двухуровневая иерархическая нейронная сеть с механизмом внимания | F1-мера (General Psychotherapy Corpus), | 71,6%, |
| | | | F1-мера (DAIC-WOZ), | 70,3%, |
| | | | UAR (DAIC-WOZ) | 70,3% |
| Huang Zh. et al. [41] | Аудио | FVTC-CNN | UAR (SH2-FS), | 68,0%, |
| | | | UAR (DAIC-WOZ) | 88,0% |
| Zhao Z. et al. [42] | Аудио | Гибридная сеть (сеть с механизмом внутреннего внимания, глубокая сверточная сеть, SVR) | MAE (Corpus 2013), | 9,65, |
| | | | RMSE (Corpus 2013), | 7,38, |
| | | | MAE (Corpus 2014), | 9,57, |
| | | | RMSE (Corpus 2014) | 7,94 |
| Seneviratne N. et al. [43] | Аудио | SVM | Accuracy | 81,7% |
| Stankevich N. et al. [44] | Текст | Первая модель; Вторая модель | F1-мера, | 63,0%, |
| | | | Precision, | 65,0%, |
| | | | Recall; | 61,0%; |
| | | | Recall, | 84,0%, |
| | | | F1-мера | 61,0% |
| Enikolopov S.N. et al. [45] | Текст | Метод случайного леса, психолингвистические признаки и биграммы | F1-мера | 73,0% |

В работе [37] в качестве данных для обучения использовались два свободно доступных набора данных: AVEC2013 и AVEC2014. После вычисления и сегментирования MFCC авторы использовали lp-нормированное сжатие, объединенное с LASSO (Least Absolute Shrinkage and Selection Operator), чтобы найти оптимальный параметр сжатия с целью последующей генерации признаков на уровне высказываний для определения депрессии. Эти данные использовались для обучения гибридной модели, которая содержит CNN, LSTM и DNN, а итоговое предсказание уровня депрессии проводилось с использованием SVR (Support Vector Regression). В результате экспериментов было выявлено, что lp-нормированное сжатие с LASSO (где параметр p = 4,06 в AVEC2013 и p = 2,13 в AVEC2014) оказалось наилучшим вариантом сжатия. На тестовом наборе AVEC2013 были получены результаты RMSE = 9,79, MAE = 7,48, а на тестовом наборе AVEC2014 – RMSE = 9,66, MAE = 8,02.

Авторы работы [38] для обучения моделей использовали многомодальный корпус DAIC-WOZ. Для данного исследования авторы решили прибегнуть к искусственному выравниванию между текстовыми, аудио- и видеопризнаками, чтобы получить точные временные метки каждого произнесенного слова. На каждой временной отметке они выравнивали слова и соответствующие им отрезки аудио с использованием инструментария Penn Phonetics Lab Forced Aligner (P2FA), который может применяться для сравнения транскрипций с аудиофайлами, фонема за фонемой. Путем ручной проверки искусственное выравнивание было проделано с достаточно высокой точностью для изучения объединения модальностей. В модели использовались highway слои (имплементация стохастического градиентного спуска) прямого распространения с оконным механизмом, которые обучаются регулировать поток информации в сети, присваивая веса видео и аудио входным данным на каждой временной отметке. Результаты обучения LSTM с механизмом окна на признаках трех модальностей: F1-мера = 81,0%, Precision = 80,0%, MAE = 3,61, RMSE = 4,99.

В работе [39] для обучения системы использовали записи 110 человек, из которых 54 никогда не имели психических заболеваний и были обозначены как контрольная группа, а у 56 была диагностирована депрессия, они были определены в группу людей с депрессией. В группе информантов с депрессией у информантов имелись следующие заболевания: большое депрессивное расстройство (19 случаев), биполярное расстройство в депрессивной фазе или с последним депрессивным эпизодом (13 случаев), реактивная депрессия (7 случав), эндо-реактивная депрессия (6 случаев) и тревожно-депрессивное расстройство (4 случая). Для остальных 7 информантов точный диагноз не был установлен.

Все участники являлись носителями итальянского языка, их просили прочитать вслух басню Эзопа «Ветер и Солнце». В качестве акустических признаков использовался набор Interspeech 2009 Emotion Challenge, который был расширен путем добавления признаков для определения скорости чтения и использования пауз. Обучение классификатора SVM, имплементированного в библиотеке Scikit-learn, происходило при помощи техники leave-one-out. Авторы предположили, что люди с депрессией читают медленнее и чаще используют длинные паузы. Данное предположение подкрепляется исследованиями нейробиологов, которые показывают, что течение процессов в мозге, связанных с языком, занимают больше времени у людей с депрессией. В частности, было показано, что есть связь между депрессией и дисфункцией в некоторых зонах, участвующих в семантической обработке языка, включая фронтальную извилину и префронтальный кортекс. В проведенных экспериментах среднее время на чтение текста и стандартное отклонение 54,92 +/- 2,66 сек и 47,38 +/- 1,20 сек для людей с депрессией и без соответственно, а скорость чтения 202,1 и 234,3 слова в минуту. Так, после добавления этих признаков, авторам удалось улучшить точность распознавания депрессии с 68,2% до 84,5%.

В работе [40] для обучения использовались корпусы General Psychotherapy Corpus и DAIC-WoZ. Авторы предложили подход с использованием иерархической архитектуры нейронной сети с вниманием для определения депрессии по транскрипциям клинических интервью. Авторы предположили, что эмоциональное содержание может быть отличительным фактором между языками людей с депрессией и без. Основываясь на этом, они применили внешние лингвистические знания об эмоциональном содержимом слов, рассмотрев эмоции, тональность, валентность и психолингвистическую аннотацию для слов. Для того чтобы исследовать использование слов, которые отражают позитивную и негативную тональности, грусть и тревожность, авторы использовали инструментарий LIWC lexicon, в котором представлена психолингвистическая аннотация 18504 слов для 73 различных категорий слов. Эксперименты показали, что дополнительная информация об эмоциях улучшает результат предложенной архитектуры. Авторам удалось добиться F1-меры 71,6% для корпуса General Psychotherapy Corpus, а также F1-меры и невзвешенной средней полноты 70,3%: для корпуса DAIC-WOZ.

Работа [41] посвящена проблеме обобщаемости и предлагает несколько стратегий адаптации, которые модернизируют предобученные модели на основе расширяемых сверточных сетей с целью улучшить точность определения депрессии как в лабораторных, так и в естественных

условиях. Для обучения сетей использовались два корпуса: SH2-FS (Free Speech) и DAIC-WOZ. Авторы использовали четыре набора акустических признаков: 3 форманты, 13 спектральных центроидных частот, 16 MFCC и 16 дельт MFCC. В работе исследуется метод FVTC-CNN (Full Vocal Tract Coordination – Convolutional Neural Networks). Он состоит из двух частей: матрицы по типу изображений FVTC и расширяемых CNN. Авторам удалось добиться точности по невзвешенной средней полноте 68,0% для корпуса SH2-FS и 88,0% для корпуса DAIC-WOZ.

В работе [42] изучались преимущества гибридной сети, которая кодирует характеристики речи, относящиеся к депрессии. В работе использовался корпус AViD-Corpus, представленный на соревнованиях по аудиовизуальному определению эмоций в 2013 и 2014 годах. Для вычисления низкоуровневых признаков использовался набор eGeMAPS из openSMILE. Предложенный метод включает в себя сети внутреннего внимания, обученные на низкоуровневых акустических признаках, глубокую сверточную сеть, обученную информации из трехмерных логмел спектрограмм, и модуль предсказания степени депрессии. В сверточной сети применяется сжатие по среднему для объединения дополнительных признаков на уровне высказываний, которые являются входными данными для регрессора опорных векторов, который в итоге предсказывает баллы по шкале Бека-2. Для корпуса 2013 года были получены результаты 9,65 и 7,38 по показателям RMSE и MAE соответственно. Для корпуса 2014 года авторам удалось добиться результатов RMSE = 9,57 и MAE = 7,94.

Авторы работы [43] исследовали изменения в речи, которые происходят в результате психомоторной заторможенности, считающейся ключевым признаком БДР. Для этого применялись инверсированные переменные речевого тракта, которые получаются путем использования системы инверсии речи, преобразующей акустический сигнал в шесть артикуляторных траекторий. Также были использованы мел-частотные кепстральные коэффициенты и корреляционные признаки. В качестве данных использовался корпус Mundt, а классификатор SVM обучался с использованием LOSO. Авторам удалось добиться точности (Accuracy) определения депрессии 81,7%.

В цикле работ [44–48] авторы используют текстовую модальность для определения депрессии, а именно частеречный анализ, TF-IDF, векторные преставления слов, n-граммы, классические текстовые и психолингвистические признаки, а также анализ тональности. Наилучший результат был получен авторами на основе набора данных CLEF/eRisk 2017, в который входит коллекция текстовых сообщений 887 пользователей социальной сети Reddit, из которых 135 текстов помечены как депрессивные.

Использовались методы SVM и случайного леса, имплементированные в библиотеке Scikit-learn. Лучшие результаты определения депрессии были получены при использовании SVM и модели TF-IDF со стилистическими и морфологическими признаками, 63,0% по показателю F1-меры, 61,0% полноты и 65,0% точности. При этом модель на основе SVM, векторного представления слов и стилистических признаков получила наилучший результат полноты, равный 84,0%, и F1-меры, равной 61,0%. При экспериментах со случайным лесом наилучшей моделью оказалась TF-IDF с морфологическими признаками, при помощи которой был получен результат 79,0% точности и 62,0% F1-меры. Лучший результат на корпусах, собранных авторами, достиг F1-меры в 73,0% при использовании метода случайного леса и набора признаков, включающего психолингвистические признаки и биграммы. Кроме того, авторы собрали два корпуса: корпус эссе «Я, другие, мир» и корпус информации из профилей социальной сети «Вконтакте».

Стоит отметить, что в последние несколько лет актуальными являются системы, использующие различные архитектуры нейронных сетей, а наилучшие результаты были получены при использовании рекуррентных архитектур и архитектур с механизмом внимания. Данная особенность присуща большинству систем, представленных как на соревнованиях AVEC, так и вне соревнований. Описанные работы показывают эффективность многомодального подхода при определении депрессии. Также, нельзя не заметить, что вербальная информация играет важную роль и позволяет добиться высоких результатов, наряду с паттернами невербальной информации, что подтверждает опыт терапевтов.

**6. Систематизация современных методов, информативных признаков и классификаторов, используемых при разработке многомодальных систем автоматического определения депрессии.** В ходе проведенного аналитического обзора был выявлен ряд сложностей при автоматическом определении депрессии: 1) каждый поведенческий сигнал предоставляет только частичную информацию, которая может быть комбинированной формой более реалистичной модели определения поведенческих индикаторов депрессии; 2) некоторая полезная информация может быть недоступна или в принципе скрыта; 3) определение базового поведения может быть осложнено ввиду ограничений в поведенческих данных.

Поскольку сбор данных для обучения моделей в системах автоматического определения депрессии является трудоемким и сложным из-за специфики задачи, существующие на данный момент корпусы имеют относительно небольшое количество данных, а также имеют дисбаланс в

количестве экземпляров в классах обучающих данных. Для решения такой проблемы возможно использование различных показателей точности работы моделей, учитывающих дисбаланс в количестве экземпляров в классах, техник аугментации данных и выбора наиболее информативных признаков. В описанных выше работах для выбора наиболее информативных признаков использовались следующие методы: выбор признаков на основе корреляции, метод главных компонент и метод частных наименьших квадратов.

Для вычисления признаков каждой модальности в рассмотренных работах использовались методы, которые можно классифицировать по трем группам: аудио, видео, текст, среди которых существует разделение на нейросетевые методы и программные инструментарии, они представлены на рисунке 2. Также на нем представлены основные информативные признаки депрессии по аудиосигналу, видеоряду и лексике, которые можно получить при использовании методов вычисления признаков.

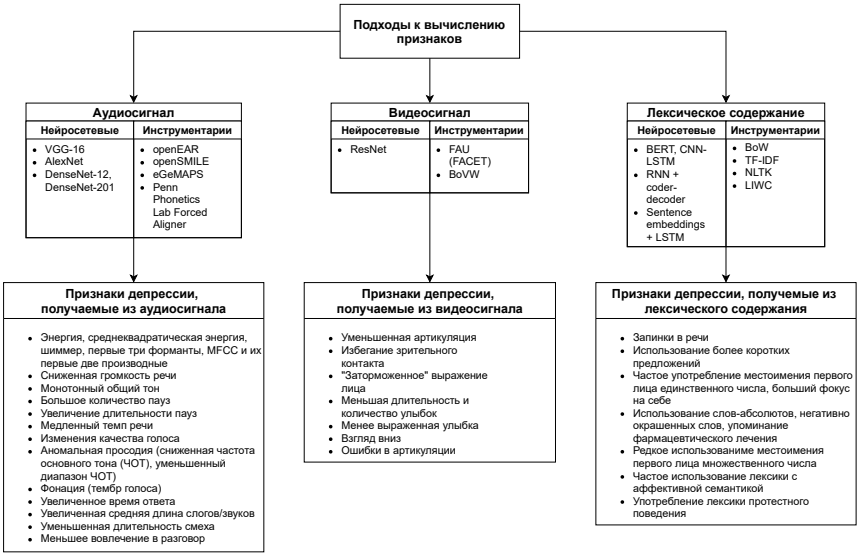


Рис. 2. Систематизация методов для вычисления информативных признаков и основные признаки депрессии, которые используются в современных автоматических системах

Наблюдаются изменения в речеобразовании у людей с депрессией после лечения, заключающиеся в изменениях тона голоса, громкости, частоты, артикуляции, беглости речи. В исследовании [49] показано, что

психологические особенности человека влияют на особенности написанного им текста. Наиболее чувствительным к психологическим особенностям человека оказался показатель частоты лексики с аффективной семантикой. Авторы выявили, что при высоких показателях депрессивности и тревожности, чувстве собственной незначительности и сниженной стратегией самоконтроля чаще употребляется лексика протестного поведения.

Среди рассмотренных работ можно выделить регрессионные и классификационные системы, в которых используются как нейросетевые, так и классические классификаторы, они представлены на рисунке 3. Так, можно отметить, что для задач классификации и регрессии при определении депрессии популярны в основном нейросетевые методы, а именно сложные архитектуры нейросетевых методов. Вероятно, данная тенденция прослеживается ввиду того, что такие методы обладают большей устойчивостью к переобучению, большей способностью к обобщению, а также способностью к выявлению скрытых корреляций в признаковом пространстве.

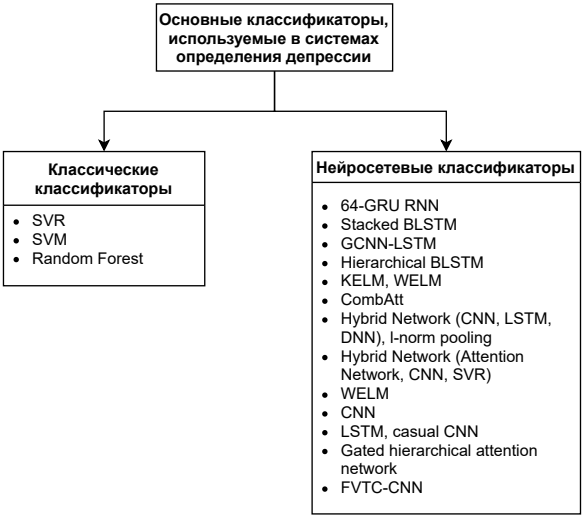


Рис. 3. Систематизация классификаторов, используемых в рассмотренных автоматических системах определения депрессии

На основе проведенного анализа можно сформулировать потенциальные требования, которые могли бы быть выдвинуты к разрабатываемым автоматическим системам определения депрессии, а именно:

1. Использование максимально возможного количества модальностей ввиду того, что специалистами учитываются все модальности при личной беседе. Кроме того, такой подход позволяет расширить возможности применения автоматических систем, так как будет возможность анализа дополнительных паралингвистических явлений.

2. Результат верного распознавания депрессии должен быть как можно выше (на данный момент лучший показатель точности среди автоматических систем находится на уровне 81% по показателю F-меры), так как сфера медицины относится к тем сферам применения, где ложные срабатывания автоматической системы могут быть критическими.

3. Для качественного выполнения второго пункта требований автоматические системы должны проходить тестирование в максимально приближенных к реальной жизни условиях.

4. Апробация в реальной жизни на этапе тестирования должна проходить под контролем специалистов, которые могли бы подтвердить верные предсказания или скорректировать ложные.

**7. Заключение.** Представлен аналитический обзор научных исследований за последние два года, посвященных разработке автоматических систем определения депрессивного состояния у людей. По данным ВОЗ, одним из наиболее распространенных психических расстройств является депрессия. Количество работ по автоматическому определению депрессии подтверждает возросший в последние годы интерес к теме, а также ее актуальность, поскольку депрессивное состояние распространено повсеместно и имеет свойство приводить к ухудшению жизнедеятельности человека и даже инвалидности или смерти. На данный момент используется множество признаков и показателей для объективной диагностики депрессивного состояния. Кроме того, существуют различия по вариантам течения и тяжести заболевания, подтипам. Специалисты определяют наличие депрессии путем беседы и заполнения различного рода опросников, однако такая оценка может варьироваться в зависимости от множества факторов, а потому на данный момент не существует объективного метода диагностики депрессии.

Рассмотрены теоретические и практические работы, представленные как на соревнованиях по аудиовизуальному распознаванию эмоций и определению депрессии, так и вне соревнований. Согласно теоретическим работам, депрессия непосредственно относится к расстройствам настроения, а значит, касается и эмоций в том числе. Это подтверждается тем, что существует сильная корреляция с аффективными величинами валентности, активации и доминации. При этом наблюдается обратная корреляция между депрессией и агрессией. Работы, в которых прово-

дилось межкультурное сравнение симптомов депрессии, показали, что сочетания признаков депрессии обнаруживаются во всех культурах. Из этого следует, что автоматические системы определения депрессии могут быть универсальны. В практических работах были представлены одномодальные и многомодальные системы за последние два года, которые решали как задачу регрессивной классификации для определения степени тяжести депрессии, так и задачу бинарной классификации наличия заболевания или его отсутствия. Для создания автоматической системы определения депрессии в рассмотренных работах были предложены различные подходы обработки многомодальных данных и построения моделей машинного обучения. Так, в такой системе могут применяться как линейные методы, так и сложные нейросетевые методы для обучения модели. Также в работе перечислены инструментарии и нейросетевые методы для вычисления признаков в различных модальностях.

В результате выполненного аналитического обзора можно сделать вывод, что объединение модальностей позволяет улучшить результаты определения депрессии (функционирование в сложных условиях при зашумленной или неразборчивой речи, при отсутствии речи и пр.), кроме того, анализ нескольких модальностей снижает вариативность при анализе речи и может значительно улучшить качество распознавания, поскольку анализируются дополнительные паралингвистические аспекты (движения бровей, напряжение губ, направление взгляда, движения рук и др.), которые могут являться информативными характеристиками. Также анализ лексической составляющей речевого высказывания может позволить обнаружить различные показатели в речи, например, неуверенность, выражающуюся в отдельных словах, паузах хезитации и междометиях и другом. В дальнейших работах мы планируем разработать собственный прототип автоматической системы определения депрессии по речи и исследовать на практике признаки депрессии, выявленные в ходе данного аналитического обзора, как со стороны акустических, так и со стороны лексических характеристик.

## Литература

1. World Health Organization. 2017. Depression and Other Common Mental Disorders: Global Health Estimates. Technical Report. World Health Organization. Licence: CC BY-NC-SA 3.0 IGO.
2. GBD 2017 Disease and Injury Incidence and Prevalence Collaborators. (2018). Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. The Lancet. DOI:https://doi.org/10.1016/S0140-6736(18)32279-7.
3. *Spitzer R.L.* Patient health questionnaire: PHQ. New York State Psychiatric Institute. 1999.

4. *Beck A.T., Ward C.H., Mock J., et al.* An inventory for measuring depression. Archives of General Psychiatry. 1961. vol. 4. pp. 561–571. DOI:https://doi:10.1001/archpsyc.1961.01710120031004.
5. *Rush A.J., Trivedi M.H., Ibrahim H.M., et al.* The 16-item Quick Inventory of Depressive Symptomatology (QIDS), clinician rating (QIDS-C), and self-report (QIDS-SR): A psychometric evaluation in patients with chronic major depression. Biological Psychiatry. 2003. vol. 54. no.5. pp. 573–583. DOI:https://doi.org/10.1016/S0006-3223(02)01866-8.
6. *Gonzalez J.S., Shreck E., Batchelder A.* Hamilton Rating Scale for Depression (HAM-D). In: Gellman MD, Turner JR, editors. Encyclopedia of behavioral medicine. New York: Springer. 2013. pp. 887–888. DOI:https://doi.org/10.1007/978-1-4419-1005-9_198.
7. *Valstar M., Schuller B., Smith K., et al.* AVEC 2013: the continuous audio/visual emotion and depression recognition challenge. Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge (AVEC'13). Association for Computing Machinery, New York, NY, USA. 2013. pp. 3–10. DOI:https://doi.org/10.1145/2512530.2512533.
8. *Valstar M., Schuller B., Smith K., et al.* AVEC 2014 — 3D dimensional affect and depression recognition challenge. Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge, Workshop of MM. 2014. pp. 3-10. DOI:https://doi.org/10.1145/2661806.2661807.
9. *Valstar M., Gratch J., Schuller B., et al.* Summary for AVEC 2016: Depression, Mood, and Emotion Recognition Workshop and Challenge. Proceedings of the 24th ACM international conference on Multimedia (MM '16). Association for Computing Machinery, New York, NY, USA. 2016. pp. 1483–1484. DOI:https://doi.org/10.1145/2964284.2980532.
10. *Ringeval F., Schuller B., Valstar M., et al.* AVEC 2017: Real-life Depression, and Affect Recognition Workshop and Challenge. Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge (AVEC '17). Association for Computing Machinery, New York, NY, USA. 2017. pp. 3–9. DOI:https://doi.org/10.1145/3133944.3133953.
11. *Ringeval F., Schuller B., Valstar M., et al.* AVEC 2019 Workshop and Challenge: State-of-Mind, Detecting Depression with AI, and Cross-Cultural Affect Recognition. In Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop (AVEC '19). Association for Computing Machinery, New York, NY, USA. 2019. pp. 3–12. DOI:https://doi.org/10.1145/3347320.3357688.
12. *Потапова Р.К.* Вариативность акустических параметров звучащей речи. Вестник Московского государственного лингвистического университета. Гуманитарные науки. Гуманитарные науки. 2016. т. 740. с. 137-147.
13. *Stahl S.M.* Stahl's essential psychopharmacology: Neuroscientific basis and practical applications. Cambridge: Cambridge University Press (4th ed.). 2013. P. 628.
14. American Psychiatric Association. Diagnostic and statistical manual of mental disorders (5th ed.). 2013. P. 992. DOI:https://doi.org/10.1176/appi.books.9780890425596.
15. *Franklin J.C., Ribeiro J.D., Fox K.R., et al.* Risk factors for suicidal thoughts and behaviors: a meta-analysis of 50 years of research. Psychol Bull. 2017. vol. 143. no. 2. pp. 187-232. DOI:https://doi.org/10.1037/bul0000084.
16. *Belsher B.E., Smolenski D.J., Pruitt L.D., et al.* Prediction Models for Suicide Attempts and Deaths: A Systematic Review and Simulation. JAMA Psychiatry. 2017. vol. 76. no. 6. pp. 642–651.
17. *Singer K.* Depressive disorders from a transcultural perspective. Social Science & Medicine. 1975. vol. 9. 289-301. DOI:https://doi.org/10.1016/0037-7856(75)90001-3.
18. *Beck A.T., Steer R.A., Brown G.* Beck Depression Inventory–II. APA PsycTests. 1996. P.38. DOI:https://doi.org/10.1037/t00742-000.

19. *Alghowinem S., Goecke R., Wagner M., et al.* From joyous to clinically depressed: Mood detection using spontaneous speech. Proceedings of FLAIRS Conference, G. M. Youngblood and P. M. McCarthy, Eds. AAAI Press. 2012. pp. 141–146.
20. *Yang Y., Fairbairn C., Cohn J.* Detecting depression severity from vocal prosody. IEEE Transactions on Affective Computing. 2013. vol. 4. no. 2. pp. 142–150.
21. *Gratch J., et al.* The Distress Analysis Interview Corpus of Human and Computer Interviews. Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland. 2014. pp. 3123-3128.
22. *Litvinova T., Ryzhkova E., Litvinova O.* Features of Written Texts of People with Different Profiles of the Lateral Brain Organization of Functions (on the Basis of RusNeuroPsych Corpus). Proceedings of 7th Tutorial and Research Workshop on Experimental Linguistics, ExLing 2016, International Speech Communication Association, Saint Peters-burg, Russia. 2016. pp. 107-110.
23. *Mundt J.C., Snyder P.J., Cannizzaro M.S., et al.* Voice acoustic measures of depression severity and treatment response collected via interactive voice response (ivr) technology. Journal of Neurolinguistics. 2007. vol. 20. no. 1. pp. 50 – 64.
24. General Psychotherapy Corpus. URL: http://alexanderstreet.com. (дата обращения: 10.12.2020).
25. *Huang Z., Epps J., Joachim D., et al.* Depression detection from short utterances via diverse smartphones in natural environmental conditions. Proceedings of Interspeech. 2018. pp. 3393–3397.
26. *Willmott C.J., Matsuura K.* Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Climate Research. 2005. vol. 30. pp. 79–82. DOI:https://doi.org/10.3354/cr030079.
27. *Lin L.I.* A concordance correlation coefficient to evaluate reproducibility. Biometrics. 1989. vol 45. no. 1. pp. 255-268.
28. *Altman D.G., Bland J.M.* Diagnostic tests. 1: Sensitivity and specificity. BMJ (Clinical research ed.). 1994. vol. 308. no. 6943. P. 1552. DOI:https://doi.org/10.1136/bmj.308.6943.1552.
29. *Ringeval F., Schuller B., Valstar M., et al.* AVEC 2019 Workshop and Challenge: State-of-Mind, Detecting Depression with AI, and Cross-Cultural Affect Recognition. In Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop (AVEC '19). Association for Computing Machinery, New York, NY, USA. 2019. pp. 3–12. DOI:https://doi.org/10.1145/3347320.3357688.
30. *Kaya H., Fedotov D., Dresvyanskiy D., et al.* Predicting depression and emotions in the crossroads of cultures, paralinguistics, and non-linguistics. Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop (AVEC '19). Association for Computing Machinery, New York, NY, USA. 2019. pp. 27–35. DOI:https://doi.org/10.1145/3347320.3357691.
31. *Ray A., Kumar S., Reddy R., et al.* Multi-level Attention Network using Text, Audio and Video for Depression Prediction. In Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop (AVEC '19). Association for Computing Machinery, New York, NY, USA. 2019. pp. 81–88. DOI:https://doi.org/10.1145/3347320.3357697.
32. *Makiuchi M.R., Warnita T., Uto K., et al.* Multimodal Fusion of BERT-CNN and Gated CNN Representations for Depression Detection. In Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop (AVEC '19). Association for Computing Machinery, New York, NY, USA. 2019. pp. 55–63. DOI:https://doi.org/10.1145/3347320.3357694.
33. *Fan W., He Z., Xing X., et al.* Multi-modality Depression Detection via Multi-scale Temporal Dilated CNNs. In 9th International Audio/Visual Emotion Challenge and

Workshop (AVEC '19), Association for Computing Machinery, New York, NY, USA. 2019. pp. 73–80. DOI:https://doi.org/10.1145/3347320.3357695.

34. *Yin S., Liang X., Ding H., et al.* A Multi-Modal Hierarchical Recurrent Neural Network for Depression Detection. In 9th International Audio/Visual Emotion Challenge and Workshop (AVEC '19), Association for Computing Machinery, New York, NY, USA. 2019. pp. 65-71. DOI:https://doi.org/10.1145/3347320.3357696.
35. *Haque A., Guo M., Miner A.S., et al.* Measuring Depression Symptom Severity from Spoken Language and 3D Facial Expressions. Machine Learning for Health (ML4H) Workshop at NeurIPS 2018, Montréal, Canada. 2018. [Online]. Available: http://arxiv.org/abs/1811.0859.
36. *Qureshi S.A., Hasanuzzaman M., Saha S., et al.* The Verbal and Non Verbal Signals of Depression — Combining Acoustics, Text and Visuals for Estimating Depression Level.[Online]. Available: http://arxiv.org/abs/1904.07656.
37. *Niu M., Tao J., Liu B., et al.* Automatic Depression Level Detection via lp-Norm Pooling. Proceedings of Interspeech. 2019. pp. 4559-4563.
38. *Rohanian M., Hough J., Purver M.* Detecting depression with word-level multimodal fusion. Proceedings of Interspeech. 2019. pp. 1443-1447.
39. *Tao F., Esposito A., Vinciarelli A.* Spotting the traces of depression in read speech: An Approach Based on Computational Paralinguistics and Social Signal Processing. Proceedings of Interspeech. 2020. pp.1828-1832.
40. *Xezonaki D., Paraskevopoulos G., Potamianos A., et al.* Affective Conditioning on Hierarchical Networks applied to Depression Detection from Transcribed Clinical Interviews. Proceedings of Interspeech. 2020. pp. 4556-4560.
41. *Huang Zh., Epps J., Joachim D., et al* Domain Adaptation for Enhancing Speech-based Depression Detection in Natural Environmental Conditions Using Dilated CNNs. Proceedings of Interspeech. 2020. pp. 4561-4565.
42. *Zhao Z., Li Q., Cummins N., et al.* Hybrid Network Feature Extraction for Depression Assessment from Speech. Proceedings of Interspeech. 2020. pp. 4956-4960.
43. *Seneviratne N., Williamson J.R., Lammert A.C., et al.* Extended Study on the Use of Vocal Tract Variables to Quantify Neuromotor Coordination in Depression. Proceedings of Interspeech. 2020. pp. 4551-4555.
44. *Stankevich M., Isakov V., Devyatkin D., et al.* Feature Engineering for Depression Detection in Social Media. Proceedings of the 7th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2018). 2020. pp. 426-431.
45. *Ениколопов С.Н., Медведева Т.И., Воронцова О.Ю., и др.* Лингвистические характеристики текстов психически больных и здоровых людей. Психологические исследования. 2018. т. 11. №61. с. 1.
46. *Kuznetsova Y.M., Kiselnikova N.V., Enikolopov S.N. et al.* Predicting Depression from Essays in Russian. Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2019". 2019. pp. 647-657.
47. *Stankevich M., Smirnov I., Kiselnikova N., et al.* Depression Detection from Social Media Profiles. Data Analytics and Management in Data Intensive Domains. DAMDID/RCDL 2019. Communications in Computer and Information Science. 2019. vol. 1223. pp. 181-194.
48. *Stankevich M., Ignatiev N. Smirnov I.* Predicting Depression with Social Media Images. Proceedings of the 9th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2020). 2020. pp. 235-240.
49. *Ениколопов С.Н., Кузнецова Ю.М., Пенкина М.Ю., и др.* Особенности текста и психологические особенности: опыт эмпирического компьютерного исследования. Труды Института системного анализа РАН. 2019. т. 69. №3. С. 91-99.

**Величко Алёна Николаевна** — младший научный сотрудник/аспирант, лаборатория речевых и многомодальных интерфейсов, СПб ФИЦ РАН. Область научных интересов: машинное обучение, речевые технологии, компьютерная паралингвистика, определение деструктивных проявлений по речи. Число научных публикаций — 10. velichko.a.n@mail.ru; http://hci.nw.ru/ru; 14-я линия В.О., д. 39, г. Санкт-Петербург, 199178, РФ; р.т.: +7-(812)-328-04-21, +7-(812)-328-70-81.

**Карпов Алексей Анатольевич** — д-р техн. наук, доцент, главный научный сотрудник/руководитель, лаборатория речевых и многомодальных интерфейсов, СПб ФИЦ РАН. Область научных интересов: многомодальные интерфейсы, распознавание речи, речевые технологии, компьютерная паралингвистика. Число научных публикаций — 300+. karpov@iias.spb.su; http://hci.nw.ru/ru; 14-я линия В.О., д. 39, г. Санкт-Петербург, 199178, РФ; р.т.: +7-(812)-328-04-21, +7-(812)-328-70-81.

**Поддержка исследований.** Работа выполнена при финансовой поддержке РФФИ фонда (проект № 20-37-90144), а также частично в рамках бюджетной темы № 0073-2019-0005.

**DOI** 10.15622/ia.2021.3.1

A. VELICHKO, A. KARPOV

# ANALYTICAL REVIEW OF AUTOMATIC SYSTEMS FOR DEPRESSION DETECTION BY SPEECH

*Velichko A., Karpov A.* **Analytical Review of Automatic Systems for Depression Detection by Speech.**

**Abstract.** In recent years the interest in automatic depression detection has grown within medical and scientific-technical communities. Depression is one of the most widespread mental illnesses that affects human life. In this review we present and analyze the latest researches devoted to depression detection. Basic notions related to the definition of depression were specified, the review includes both unimodal and multimodal corpora containing records of informants diagnosed with depression and control groups of non-depressed people.

Theoretical and practical researches which present automated systems for depression detection were reviewed. The last ones include unimodal as well as multimodal systems. A part of reviewed systems addresses the challenge of regressive classification predicting the degree of depression severity (non-depressed, mild, moderate and severe), and another part solves a problem of binary classification predicting the presence of depression (if a person is depressed or not). An original classification of methods for computing of informative features for three communicative modalities (audio, video, text information) is presented. New methods for depression detection in every modality and all modalities in total are defined. The most popular methods for depression detection in reviewed studies are neural networks. The survey has shown that the main features of depression are psychomotor retardation that affects all communicative modalities and strong correlation with affective values of valency, activation and domination, also there has been observed an inverse correlation between depression and aggression. Discovered correlations confirm interrelation of affective disorders and human emotional states. The trend observed in many reviewed papers is that combining modalities improves the results of depression detection systems.

**Keywords:** Automatic Depression Detection by Speech, Computational Paralinguistics, Speech Technologies, Machine Learning

**Velichko Alena** — Junior Researcher/Ph.D., Student, Speech and Multimodal Interfaces Laboratory, SPC RAS. Research interests: machine learning, speech technologies, computational paralinguistics, detection of destructive behaviour by speech. The number of publications — 10. velichko.a.n@mail.ru; http://hci.nw.ru/ru; 39, 14-th Line V.O., St. Petersburg, 199178, Russia; office phone: +7-(812)-328-04-21, +7-(812)-328-70-81.

**Karpov Alexey** — Ph.D., Dr. Sci., Associate Professor, Head of Laboratory, Laboratory of Speech and Multimodal Interfaces, SPC RAS. Research interests: multimodal interfaces, speech recognition, speech technologies, computational paralinguistics. The number of publications —300+. karpov@iias.spb.su; http://hci.nw.ru/ru; 39, 14-th Line V.O., St. Petersburg, 199178, Russia; office phone: +7-(812)-328-04-21, +7-(812)-328-70-81.

**Acknowledgements.** This research was financially supported by RFBR (grant No. 20-37-90144), as well as partially in the framework of the state research № 0073-2019-0005.

## References

1. World Health Organization. 2017. Depression and Other Common Mental Disorders: Global Health Estimates. Technical Report. World Health Organization. Licence: CC BY-NC-SA 3.0 IGO.
2. GBD 2017 Disease and Injury Incidence and Prevalence Collaborators. (2018). Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases

and injuries for 195 countries and territories, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. The Lancet. DOI:https://doi.org/10.1016/S0140-6736(18)32279-7.

3. Spitzer R.L. Patient health questionnaire: PHQ. *New York State Psychiatric Institute*. 1999.
4. Beck A.T., Ward C.H., Mock J., et al. An inventory for measuring depression. *Archives of General Psychiatry*. 1961. vol. 4. pp. 561–571. DOI:https://doi:10.1001/archpsyc.1961.01710120031004.
5. Rush A.J., Trivedi M.H., Ibrahim H.M., et al. The 16-item Quick Inventory of Depressive Symptomatology (QIDS), clinician rating (QIDS-C), and self-report (QIDS-SR): A psychometric evaluation in patients with chronic major depression. *Biological Psychiatry*. 2003. vol. 54. no.5. pp. 573–583. DOI:https://doi.org/10.1016/S0006-3223(02)01866-8.
6. Gonzalez J.S., Shreck E., Batchelder A. Hamilton Rating Scale for Depression (HAM-D). In: Gellman MD, Turner JR, editors. *Encyclopedia of behavioral medicine*. New York: Springer. 2013. pp. 887–888. DOI:https://doi.org/10.1007/978-1-4419-1005-9_198.
7. Valstar M., Schuller B., Smith K., et al. AVEC 2013: the continuous audio/visual emotion and depression recognition challenge. Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge (AVEC'13). Association for Computing Machinery, New York, NY, USA. 2013. pp. 3–10. DOI:https://doi.org/10.1145/2512530.2512533.
8. Valstar M., Schuller B., Smith K., et al. AVEC 2014 — 3D dimensional affect and depression recognition challenge. Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge, Workshop of MM. 2014. pp. 3-10. DOI:https://doi.org/10.1145/2661806.2661807.
9. Valstar M., Gratch J., Schuller B., et al. Summary for AVEC 2016: Depression, Mood, and Emotion Recognition Workshop and Challenge. Proceedings of the 24th ACM international conference on Multimedia (MM '16). Association for Computing Machinery, New York, NY, USA. 2016. pp. 1483–1484. DOI:https://doi.org/10.1145/2964284.2980532.
10. Ringeval F., Schuller B., Valstar M., et al. AVEC 2017: Real-life Depression, and Affect Recognition Workshop and Challenge. Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge (AVEC '17). Association for Computing Machinery, New York, NY, USA. 2017. pp. 3–9. DOI:https://doi.org/10.1145/3133944.3133953.
11. Ringeval F., Schuller B., Valstar M., et al. AVEC 2019 Workshop and Challenge: State-of-Mind, Detecting Depression with AI, and Cross-Cultural Affect Recognition. In Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop (AVEC '19). Association for Computing Machinery, New York, NY, USA. 2019. pp. 3–12. DOI:https://doi.org/10.1145/3347320.3357688.
12. Potapova R.K. [Variability of acoustic parameters of sounding speech]. *Variativnost' akusticheskih parametrov zvuchashhej rechi. Vestnik Moskovskogo gosudarstvennogo lingvisticheskogo universiteta. Gumanitarnye nauki.* [Bulletin of Moscow State Linguistic University. Humanitarian sciences.]. 2016. vol. 740. pp. 137-147. (In Russ.)
13. Stahl S.M. Stahl's essential psychopharmacology: Neuroscientific basis and practical applications. *Cambridge: Cambridge University Press (4th ed.)*. 2013. P. 628.
14. American Psychiatric Association. *Diagnostic and statistical manual of mental disorders (5th ed.)*. 2013. P. 992. DOI:https://doi.org/10.1176/appi.books.9780890425596.
15. Franklin J.C., Ribeiro J.D., Fox K.R., et al. Risk factors for suicidal thoughts and behaviors: a meta-analysis of 50 years of research. *Psychol Bull*. 2017. vol. 143. no. 2. pp. 187-232. DOI:https://doi.org/10.1037/bul0000084.

16. Belsher B.E., Smolenski D.J., Pruitt L.D., et al. Prediction Models for Suicide Attempts and Deaths: A Systematic Review and Simulation. *JAMA Psychiatry*. 2017. vol. 76. no. 6. pp. 642–651.
17. Singer K. Depressive disorders from a transcultural perspective. *Social Science & Medicine*. 1975. vol. 9. 289-301. DOI:https://doi.org/10.1016/0037-7856(75)90001-3.
18. Beck A.T., Steer R.A., Brown G. Beck Depression Inventory–II. *APA PsycTests*. 1996. P.38. DOI:https://doi.org/10.1037/t00742-000.
19. Alghowinem S., Goecke R., Wagner M., et al. From joyous to clinically depressed: Mood detection using spontaneous speech. Proceedings of FLAIRS Conference, G. M. Youngblood and P. M. McCarthy, Eds. AAAI Press. 2012. pp. 141–146.
20. Yang Y., Fairbairn C., Cohn J. Detecting depression severity from vocal prosody. *IEEE Transactions on Affective Computing*. 2013. vol. 4. no. 2. pp. 142–150.
21. Gratch J., et al. The Distress Analysis Interview Corpus of Human and Computer Interviews. Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland. 2014. pp. 3123-3128.
22. Litvinova T., Ryzhkova E., Litvinova O. Features of Written Texts of People with Different Profiles of the Lateral Brain Organization of Functions (on the Basis of RusNeuroPsych Corpus). Proceedings of 7th Tutorial and Research Workshop on Experimental Linguistics, ExLing 2016, International Speech Communication Association, Saint Peters-burg, Russia. 2016. pp. 107-110.
23. Mundt J.C., Snyder P.J., Cannizzaro M.S., et al. Voice acoustic measures of depression severity and treatment response collected via interactive voice response (ivr) technology. Journal of Neurolinguistics. 2007. vol. 20. no. 1. pp. 50 – 64.
24. General Psychotherapy Corpus. URL: http://alexanderstreet.com. (дата обращения: 10.12.2020).
25. Huang Z., Epps J., Joachim D., et al. Depression detection from short utterances via diverse smartphones in natural environmental conditions. Proceedings of Interspeech. 2018. pp. 3393–3397.
26. Willmott C.J., Matsuura K. Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research*. 2005. vol. 30. pp. 79–82. DOI:https://doi.org/10.3354/cr030079.
27. Lin L.I. A concordance correlation coefficient to evaluate reproducibility. *Biometrics*. 1989. vol 45. no. 1. pp. 255-268.
28. Altman D.G., Bland J.M. Diagnostic tests. 1: Sensitivity and specificity. *BMJ (Clinical research ed.)*. 1994. vol. 308. no. 6943. P. 1552. DOI:https://doi.org/10.1136/bmj.308.6943.1552.
29. Ringeval F., Schuller B., Valstar M., et al. AVEC 2019 Workshop and Challenge: State-of-Mind, Detecting Depression with AI, and Cross-Cultural Affect Recognition. In Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop (AVEC '19). Association for Computing Machinery, New York, NY, USA. 2019. pp. 3–12. DOI:https://doi.org/10.1145/3347320.3357688.
30. Kaya H., Fedotov D., Dresvyanskiy D., et al. Predicting depression and emotions in the crossroads of cultures, paralinguistics, and non-linguistics. Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop (AVEC '19). Association for Computing Machinery, New York, NY, USA. 2019. pp. 27–35. DOI:https://doi.org/10.1145/3347320.3357691.
31. Ray A., Kumar S., Reddy R., et al. Multi-level Attention Network using Text, Audio and Video for Depression Prediction. In Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop (AVEC '19). Association for Computing Machinery, New York, NY, USA. 2019. pp. 81–88. DOI:https://doi.org/10.1145/3347320.3357697.

32. Makiuchi M.R., Warnita T., Uto K., et al. Multimodal Fusion of BERT-CNN and Gated CNN Representations for Depression Detection. In Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop (AVEC '19). Association for Computing Machinery, New York, NY, USA. 2019. pp. 55–63. DOI:https://doi.org/10.1145/3347320.3357694.
33. Fan W., He Z., Xing X., et al. Multi-modality Depression Detection via Multi-scale Temporal Dilated CNNs. In 9th International Audio/Visual Emotion Challenge and Workshop (AVEC '19), Association for Computing Machinery, New York, NY, USA. 2019. pp. 73–80. DOI:https://doi.org/10.1145/3347320.3357695.
34. Yin S., Liang X., Ding H., et al. A Multi-Modal Hierarchical Recurrent Neural Network for Depression Detection. In 9th International Audio/Visual Emotion Challenge and Workshop (AVEC '19), Association for Computing Machinery, New York, NY, USA. 2019. pp. 65-71. DOI:https://doi.org/10.1145/3347320.3357696.
35. Haque A., Guo M., Miner A.S., et al. Measuring Depression Symptom Severity from Spoken Language and 3D Facial Expressions. Machine Learning for Health (ML4H) Workshop at NeurIPS 2018, Montréal, Canada. 2018. [Online]. Available: http://arxiv.org/abs/1811.0859.
36. Qureshi S.A., Hasanuzzaman M., Saha S., et al. The Verbal and Non Verbal Signals of Depression — Combining Acoustics, Text and Visuals for Estimating Depression Level. [Online]. Available: http://arxiv.org/abs/1904.07656.
37. Niu M., Tao J., Liu B., et al. Automatic Depression Level Detection via lp-Norm Pooling. Proceedings of Interspeech. 2019. pp. 4559-4563.
38. Rohanian M., Hough J., Purver M. Detecting depression with word-level multimodal fusion. Proceedings of Interspeech. 2019. pp. 1443-1447.
39. Tao F., Esposito A., Vinciarelli A. Spotting the traces of depression in read speech: An Approach Based on Computational Paralinguistics and Social Signal Processing. Proceedings of Interspeech. 2020. pp.1828-1832.
40. Xezonaki D., Paraskevopoulos G., Potamianos A., et al. Affective Conditioning on Hierarchical Networks applied to Depression Detection from Transcribed Clinical Interviews. Proceedings of Interspeech. 2020. pp. 4556-4560.
41. Huang Zh., Epps J., Joachim D., et al Domain Adaptation for Enhancing Speech-based Depression Detection in Natural Environmental Conditions Using Dilated CNNs. Proceedings of Interspeech. 2020. pp. 4561-4565.
42. Zhao Z., Li Q., Cummins N., et al. Hybrid Network Feature Extraction for Depression Assessment from Speech. Proceedings of Interspeech. 2020. pp. 4956-4960.
43. Seneviratne N., Williamson J.R., Lammert A.C., et al. Extended Study on the Use of Vocal Tract Variables to Quantify Neuromotor Coordination in Depression. Proceedings of Interspeech. 2020. pp. 4551-4555.
44. Stankevich M., Isakov V., Devyatkin D., et al. Feature Engineering for Depression Detection in Social Media. Proceedings of the 7th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2018). 2020. pp. 426-431.
45. Enikolopov S.N., Medvedeva T.I., Voroncova O. Ju., et al. [Linguistic characteristics of texts written by mentally ill and healthy people]. *Lingvisticheskie harakteristiki tekstov psihicheski bol'nyh i zdorovyh ljudej.Psihologicheskie issledovanija.* [Psychological investigations]. 2018. vol. 11. no, 61. pp. 1. (In Russ.).
46. Kuznetsova Y.M., Kiselnikova N.V., Enikolopov S.N. et al. Predicting Depression from Essays in Russian. Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2019". 2019. pp. 647-657.
47. Stankevich M., Smirnov I., Kiselnikova N., et al. Depression Detection from Social Media Profiles. Data Analytics and Management in Data Intensive Domains. DAMDID/RCDL

2019. Communications in Computer and Information Science. 2019. vol. 1223. pp. 181-194.

48. Stankevich M., Ignatiev N. Smirnov I. Predicting Depression with Social Media Images. Proceedings of the 9th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2020). 2020. pp. 235-240.

49. Enikolopov S.N., Kuznecova Ju.M., et al. [Characteristics of text and psychological characteristics: experience of empiric computational research]. *Osobennosti teksta i psihologicheskie osobennosti: opyt jempiricheskogo komp'juternogo issledovanija. Trudy Instituta sistemnogo analiza RAN.* [Proceedings of the institute of system analysis of RAS]. 2019. vol. 69. no. 3. pp. 91-99. (In Russ.).