УДК 004.934 **DOI** 10.15622/ia.24.6.6

М.Д. Долгушин, А.А. Карпов

АНАЛИТИЧЕСКИЙ ОБЗОР РЕЧЕВЫХ И МНОГОМОДАЛЬНЫХ МЕТОДОВ РАСПОЗНАВАНИЯ КОГНИТИВНЫХ НАРУШЕНИЙ ЛЮДЕЙ

Долгушин М.Д., Карпов А.А. **Аналитический обзор речевых и многомодальных методов распознавания когнитивных нарушений людей.**

Аннотация. В течение последнего десятилетия наблюдается рост количества публикаций научно-технической и медицинской направленности, посвящённых автоматическому определению на основании речевого и визуального материала таких когнитивных нарушений людей, которые возникают при таких нейродегенеративных заболеваниях, как деменция, болезнь Альцгеймера, болезнь Паркинсона и другие нарушения. Хотя данные заболевания имеют высокую степень распространения и являются одной из основных причин смертности и ранней инвалидизации людей, способов их эффективного лечения на данном этапе развития медицины отсутствуют или ограничены. В связи с этим ранняя диагностика симптомов и их облегчение вызывают значительный интерес исследователей. В фокусе современных исследований находится задача разработки автоматизированных и автоматических систем, основанных на количественных объективных методах, нейросетевых подходах, использующих различные модальности и их объединение, а также на методах интерпретируемого искусственного интеллекта. В статье представлен обзор и анализ основных исследований с 2022 года, посвящённых автоматическому одно- и многомодальному определению когнитивных нарушений людей. Представлены наиболее часто применяющиеся многомодальные корпусы, применяемые в данной задаче, такие как ADReSS, ADReSSo, TAUKADIAL и др. Описаны современные методы, используемые для выявления когнитивных нарушений на основе данных различных модальностей, представленные на международном соревновании TAUKADIAL-2024, и вне соревнований. По результатам соревнований, лучшими методами распознавания когнитивных нарушений являются ансамбли вероятностных моделей, обученные на интерпретируемых экспертных признаках и нейросетевых признаках текстов и аудио. Рассмотрены многомодальные подходы, использующие визуальную модальность для обучения глубоких нейросетевых моделей. Рассмотрено новое направление, изучающее применимость больших языковых моделей к анализу медицинских текстов и интерпретируемому предсказанию заболевания. Проведена систематизация методов извлечения информативных признаков и применяемых классификаторов. По результатам проведенного обзора сформулированы требования к системам автоматизированного определения когнитивных нарушений.

Ключевые слова: автоматическое определение когнитивных нарушений, речевые технологии в обеспечении здравоохранения, интерпретируемый искусственный интеллект, машинное обучение.

1. Введение. Сфера интеллектуальных речевых технологий в обеспечении здравоохранения и благополучия людей (healthcare and wellbeing) возникла сравнительно недавно как часть более широкой области применения методов искусственного интеллекта в медицинской сфере, но она уже представляет одно из самых

стремительно развивающихся междисциплинарных направлений науки.

В настоящее время проблема раннего медицинского определения когнитивных нарушений становится особенно актуальной. По оценкам Всемирной организации здравоохранения (ВОЗ) за 2021 год[1], деменцией, крайней стадией развития когнитивных нарушений, страдают более 55 миллионов человек (8,1% женщин и 5,4% мужчин старше 65 лет), а болезнью Паркинсона – свыше 8,5 миллионов человек. Кроме того, специалисты ВОЗ прогнозируют рост данных показателей в ближайшие годы. Методы эффективного лечения данных заболеваний на данный момент отсутствуют, однако своевременное определение заболевания и раннее начало терапии способны облегчить симптомы и улучшить качество жизни пациентов.

Несмотря на большой интерес научных сообществ, многие задачи в области определения когнитивных нарушений до сих пор остаются не решенными в силу значительной сложности точной диагностики [1]. Специалисты определяют наличие когнитивных нарушений путем различных опросников и тестирований в формате интервью, но эффективность многих из них зависит от множества факторов, поэтому без использования инвазивных средств точное обнаружение заболевания крайне затруднено.

Активное применение в исследованиях по автоматическому определению когнитивных нарушений находят многомодальные методы [2], то есть использующие несколько модальностей одновременно, в том числе естественную речь и звуки, артикуляцию лица, жесты рук и тела, направление взгляда, общее поведение, магнитно-резонансную томографию, дополнительную медицинскую информацию и т.д. Заболевания, вызывающие когнитивные нарушения, оказывают влияние на весь комплекс поведенческих признаков, но отдельные наборы данных, в основном, одномодальны [3], то есть содержат только речевой материал, видеокадры или снимки МРТ.

В данной работе решено сосредоточиться на речевых наборах, содержащих аудио или аудио и видео, поскольку разговорная речь выступает одним из индикаторов психоэмоционального состояния, например, в работах [4, 5], и, как следствие, может быть использована для автоматизированного обнаружения когнитивных отклонений. Речь многомодальна и соответственно позволяет использовать методы анализа аудио, текста и, при наличии, видео. Также данный метод имеет ряд преимуществ: неинвазивность (отсутствие необходимости

[1] https://www.who.int/ru/news/item/02-09-2021-world-failing-to-address-dementia-challenge

проникать в организм при исследовании) и доступность, связанная с широкой распространенностью средств аудио- и видеозаписи. Подобные методы позволяют, например, использовать телемедицинские технологии для исследования состояния пациента, что может быть крайне важным в случаях ограниченной мобильности, географической удалённости или по иным причинам.

Данный анализ проведен с целью продемонстрировать значимость существующих проблем в области автоматического определения когнитивных нарушений, описать используемые в настоящее время подходы и провести обзор существующих систем определения когнитивных нарушений.

В разделе 2 приводятся определения в области когнитивных расстройств и обсуждается влияние заболеваний на вербальные и невербальные аспекты речевой коммуникации. Раздел 3 посвящён описанию существующих одно- и многомодальных речевых баз данных по когнитивным нарушениям людей. Одно- и многомодальные системы, представленные в работах, опубликованных с 2022 года, описаны в разделах 4 и 5. В разделе 6 рассматриваются работы, изучающие применение больших языковых моделей для извлечения текстовых признаков и определения когнитивных нарушений. Раздел 7 содержит классификацию наиболее эффективных методов для построения автоматических систем определения когнитивных нарушений. Итоговые выводы по результатам проведённого исследования представлены в разделе 8.

2. Основные определения и проявления когнитивных нарушений в речи. В ходе определения ключевых понятий в качестве основного ориентира использованы рекомендации Министерства здравоохранения РФ. Из рекомендаций [6] следует, что когнитивное расстройство – это субъективное и/или объективно выявляемое ухудшение когнитивных функций (внимания, памяти, речи, восприятия и т.д.) по сравнению с исходным индивидуальным и/или среднестатистическим возрастным и образовательным уровнем. Распространенность умеренных когнитивных нарушений (УКН), по данным международных исследований, среди лиц старше 60 лет варьируется от 12 до 18% [7, 8]. Кроме того, распространенность УКН увеличивается с возрастом: с 10% в возрасте 70–79 лет до 25% у лиц в возрасте 80–89 лет [9]. В России масштабные эпидемиологические исследования когнитивных расстройств и деменции не проводились, однако в одном из последних исследований [10] отмечается, что среди пациентов амбулаторно-поликлинического звена 60 лет и старше

широко распространены как когнитивные расстройства, так и деменция: частота УКН составила 49,6%, деменции – 7,8%.

Крайнее проявление когнитивных нарушений – деменция. Деменция [6] определяется как нейропсихиатрический синдром, в структуре которого лежит приобретенное, длительное, клинически значительное ухудшение когнитивных функций, проявляющееся в виде тяжелых когнитивных нарушений и обусловливающее утрату привычного функционирования различной степени тяжести.

Потенциальными рисками для развития УКН и деменции являются болезнь Альцгеймера, болезнь Паркинсона, депрессия, посттравматическое стрессовое расстройство (ПТСР), хронический стресс и тревожность, боковой амиотрофический склероз, рассеянный склероз и др. [11]. По данным ВОЗ, депрессия выступает одним из наиболее распространенных психических расстройств человека. При этом она также считается одним из важных немоторных симптомов болезни Паркинсона и Альцгеймера, раннее обнаружение и своевременное лечение которых может значительно улучшить качество жизни пациентов [12].

Поскольку когнитивные расстройства оказывают разрушительное воздействие на мышление и моторику, расхождения на различных лингвистических уровнях языка и в паралингвистических проявлениях становятся заметны, что обуславливает возможность их автоматического обнаружения посредством анализа речевого сигнала [13].

На рисунке 1 представлена схема, иллюстрирующая связь некоторых нервно-психических заболеваний, вызывающих когнитивные нарушения с механизмами, через которые они оказывают влияние на речевой сигнал. Схема также сопровождена примерами акустических, лингвистических и поведенческих признаков, позволяющих обнаруживать подобные отклонения. Эта схема основана на систематизации из работы [14] и дополнена признаками визуальной модальности (мимики), рассмотренными в исследовании [15]. Более полные медицинские обзоры изменений дискурса и голосовых признаков при различных нервно-психических патологиях представлены в работах [16, 17]. Подробный обзор автоматически извлекаемых лексических и акустических признаков, применяемых для определения когнитивных нарушений и формального расстройства мышления, представлен в исследовании [18].

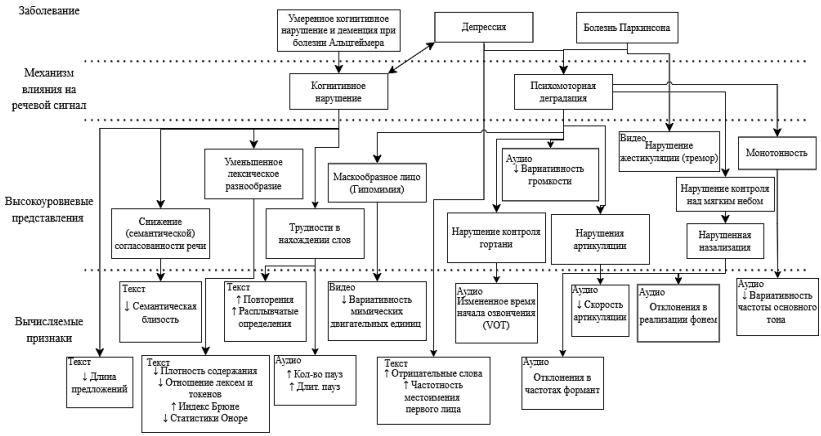

Рис. 1. Систематизация когнитивных заболеваний, нарушений и информативных признаков

Из представленной схемы видно, что влияние некоторых болезней на речевой сигнал частично пересекается. Например, депрессия и болезнь Паркинсона ассоциированы с психомоторной деградацией и, следовательно, имеют схожие проявления в речевом сигнале. Помимо этого, некоторые признаки речевых отклонений могут одновременно характеризовать различные диагнозы, поэтому использование отдельных признаков часто не является достаточным и эффективным для автоматического определения конкретных болезней. Например, депрессия может быть фактором риска для развития деменции при болезнях Альцгеймера и Паркинсона [19]. Их влияние на речевой сигнал также может быть сходным [20]. В работе [21] отмечается, что у пациентов на поздних стадиях развития болезни Альцгеймера сильно выражена апатия, что влияет на снижение общей эмоциональной выразительности мимики.

Особый интерес на схеме представляют высокоуровневые представления (согласованность, беглость речи, эмоциональное состояние и другие). С одной стороны, некоторые из этих представлений очевидны даже для неспециалистов и могут быть установлены, например, родственниками пациента. С другой стороны, большинство данных признаков, в отличие от вычисляемых признаков, не поддаются оценке с помощью количественных показателей. В связи с этим представляется перспективной задача автоматического извлечения данных признаков, например, посредством глубоких нейросетевых методов, в дополнение к уже известным

количественным признакам. Такой подход мог бы упростить интерпретацию полученных результатов, поскольку признак сам по себе был бы достаточно ясным для интерпретации.

Исходя из вышесказанного, задачи автоматического определения отдельного заболевания, являющегося причиной когнитивных нарушений, а также интерпретации автоматически обнаруженных нарушений посредством выявленных корреляций с высокоуровневыми представлениями нарушений, являются актуальными направлениями исследований.

3. Многомодальные базы данных для исследования когнитивных нарушений. Для создания автоматической системы определения возможного заболевания необходимо опираться на речевые данные как пациентов с установленным диагнозом, так и контрольной группы (КГ) без диагноза. В работах, которые рассмотрены в данном разделе, представлены одно- и многомодальные корпусы, содержащие устную речь. Большинство баз данных содержат записи пациентов с УКН и/или деменцией одного определенного типа, чаще всего – болезни Альцгеймера. Однако во многих наборах данных конкретное заболевание, являющееся причиной когнитивных нарушений, не указывается.

Во многих из представленных наборов содержится информация о показателях психологических тестов когнитивных способностей, таких как:

– Краткая шкала оценки психического статуса (Mini Mental State Examination, MMSE[2]).

– Монреальская шкала оценки когнитивных функций (Montreal Cognitive Assessment, MoCA[3]).

– Унифицированная шкала оценки болезни Паркинсона Международного общества расстройств движений (MDS-UPDRS[4]).

– Шкала депрессии А.Бека (beck depression inventory, BDI[5]).

В таблице 1 представлена характеристика наиболее распространённых наборов данных, используемых в исследованиях когнитивных нарушений людей.

DementiaBank [22] – крупнейшая опубликованная многомодальная база данных речевых записей пациентов, страдающих деменцией (в основном – УКН и БА, также присутствуют данные по

[2] https://mmse.neurol.ru/ru/index.html?ysclid=m92u2aqpfj497146286
[3] https://mocacognition.com
[4] https://mds.movementdisorders.org/UserFiles/file/MDS-UPDRS_Russian.pdf
[5] https://www.apa.org/pi/about/publications/caregivers/practice-settings/assessment/tools/beck-depression

БП и другим нарушениям). База данных содержит несколько наиболее часто встречающихся в исследованиях англоязычных корпусов и корпусы на других языках: испанский, китайский и т.д. Большинство представленных корпусов содержат аудиозаписи и текстовые транскрипции.

Pitt [23] – один из крупнейших англоязычных корпусов речевых записей пациентов с УКН и БА из базы DementiaBank. В ходе работы над созданием корпуса респондентам предлагалось выполнить задачи по описанию изображения, оценке владения языком и пересказу истории. Записи, входящие в состав корпуса, не сбалансированы по полу и возрасту, а также характеризуются различными акустическими условиями записи.

Таблица 1. Сравнение корпусов, содержащих данные людей с когнитивными нарушениями

Набор данных	Язык данных	Статистика Информантов **	Тип данных ***	Оценка	Доступность
Pitt	Англ.	104 – КГ, 208 – УКН/БА, 85 – НД	A, T, M	MMSE	По запросу [23]
ADReSS	Англ.	78 – Д, 78 – КГ.	A, T, M	MMSE	По запросу [24]
ADReSSo for AD classification	Англ.	122 – Д, 115 – КГ	A, M	MMSE	По запросу [26]
TAUKADIAL	Кит.	98 – УКН/Д, 102 – КГ	A, M	MMSE	По запросу [27]
	Англ.	123 – УКН/Д, 63 – КГ			
NCMMSC's AD dataset	Кит.	26 – БА, 44 – КГ, 54 – УКН	A, M	Метка диагноза	Открыт [28]
GAADRD	Греч.	25 – КГ, УКН и БА	A, B, M	Метка диагноза	По запросу [29]
PROMPT	Япон.	60 часов 117 КГ, Депр., БРЛ, Д	A, B, ТД	MINI, MMSE	По запросу [30]
I-CONECT	Англ.	92 – УКН, Д	A, B	MMSE	По запросу [31]
Набор ВШЭ	Русск.	100 – УКН и КГ	A, T, M	MoCA, MMSE	Закрыт [32]

* Англ. – английский, Кит. – китайский, Русск. – русский, Греч. – греческий, Япон. – японский. ** БА – болезнь Альцгеймера, БП – болезнь Паркинсона, БРЛ – биполярное расстройство личности, Д – деменция, Депр. – депрессия, КГ – контрольная группа, НД – неизвестный диагноз, РС – рассеянный склероз, УКН – умеренное когнитивное нарушение. ** A – аудио, B – видео, M – метаданные, T – Текст, ТД – трекинг движений

ADReSS (Alzheimer's Dementia Recognition through Spontaneous Speech) Challenge Dataset [24] – это стандартизированный набор данных, являющийся подкорпусом корпуса Pitt. ADReSS содержит речевые записи и транскрипции речи устных описаний изображений участниками. Аннотация была проведена с использованием системы кодирования CHAT [25]. Для всех записей была произведена очистка от шума и нормализация по громкости, поскольку многие из записей были получены в разных акустических условиях. Набор был собран с целью решения двух задач: классификации речи для определения наличия или отсутствия заболевания и регрессионной задачи предсказания результата MMSE по аудио и тексту. В 2021 была представлена расширенная версия данного набора – ADReSSo [26] (Alzheimer's Dementia Recognition through Spontaneous Speech Only), содержащая аудиоданные без ручной расшифровки, а также дополнительную выборку, которую можно использовать для предсказания когнитивного снижения с течением времени.

TAUKADIAL Challenge Dataset [27] – сбалансированный многоязычный набор данных, содержащий аудиозаписи спонтанной речи на английском и китайском когнитивно здоровых респондентов и пациентов с УКН. Набор сбалансирован по полу и возрасту с целью минимизации возможных искажений.

NCMMSC's (National Conference on Man-Machine Speech Communication) AD dataset [28] – набор данных, использованный в соревновании NCMMSC2021 AD Recognition Challenge в 2021 году. Набор содержит речь 124 носителей китайского языка, среди которых есть пациенты с когнитивными отклонениями и здоровые люди. Респонденты выполняли задания по описанию изображений и оценке владения языком, а также вели свободный диалог с интервьюером.

GAADRD (Greek Alzheimer's Association for Dementia and Related Disorders) [29] – коллекция наборов данных, собранных в рамках проекта Dem@Care, содержащих аудио, видео и физиологические показатели пациентов с диагностированной деменцией, являющихся носителями греческого языка. Записи производились в Греции в лабораторных или домашних условиях. В проекте приняли участие 25 респондентов с различной степенью тяжести когнитивных или поведенческих нарушений. Участникам предлагалось выполнять обычные повседневные дела, такие как чтение, разговор по телефону и т.д., пока их движения записывались на видео, а состояние отслеживалось с помощью различных датчиков.

PROMPT (PRoject for Objective Measures using computational Psychiatry Technology) dataset[6] [30] – аудиовизуальный речевой корпус на японском языке, содержащий записи людей старше 20 лет. Среди респондентов есть лица с депрессивными эпизодами и когнитивными нарушениями. Респонденты из контрольной группы и пациенты проходили международное нейропсихиатрическое мини-интервью (The Mini-International Neuropsychiatric Interview, M.I.N.I.) и оценивание по шкале MMSE. Видеозаписи представляют собой десятиминутные интервью респондента с медицинским специалистом на общие темы. Интервью с одними и теми же респондентами проводились до десяти раз в течение нескольких лет для отслеживания возможных изменений с течением времени. В результате суммарная длительность записей набора составила примерно 60 часов.

Internet-Based Conversational Engagement Clinical Trial (I-CONECT) data set [7] [31] – аудиовизуальный речевой корпус на английском языке, содержащий записи людей старше 75 лет с УКН. Для исследования на основании критерия качества записи были отобраны 92 видео, в которых респонденты говорят на одну из трех тем: лето, праздники и забота о себе. Набор сбалансирован, половина респондентов имеет умеренные когнитивные нарушения.

Набор ВШЭ [32] содержит 100 спонтанных рассказов, полученных от 100 русскоязычных пожилых людей с УКН и без них. В записях присутствует речь интервьюера. Записи сопровождены полной текстовой разметкой, выполненной вручную, а также разметкой речевых сбоев и пауз хезитации. Каждому респонденту сопоставлена оценка по шкалам MMSE и MoCA, утверждённая врачами-специалистами. Данный набор является одним из немногих наборов на русском языке по умеренным когнитивным нарушениям, однако он закрыт для публичного использования и крайне несбалансирован по полу и возрасту, что потенциально усложняет применение методов машинного обучения.

Существуют также наборы данных по заболеваниям, связанным с когнитивными изменениями, таким как БП, депрессия и боковой амиотрофический склероз. Это, например, открытый сбалансированный аудиовизуальный корпус In-the-wild Speech Medical Corpus (WSM) [33], содержащий подкорпусы по болезни Паркинсона и депрессии. Данный корпус был собран из видеозаписей, выложенных на сайте YouTube, и аннотирован с помощью краудсорсинга. Другой пример – PC GITA [34], сбалансированный медицинский речевой

[6] https://www.i2lab.info/
[7] https://www.i-conect.org/request-data

корпус на испанском языке, посвящённый БП и часто используемый в исследованиях, а также корпусы Distress Analysis Interview Corpus (DAIC) [35] и Extended Distress Analysis Interview Corpus (E-DAIC) [36] – многомодальные коллекции клинических интервью с пациентами с депрессией и посттравматическим синдромом.

Кроме того, следует упомянуть о наборе голосовых данных пациентов с боковым амиотрофическим склерозом [37], записанном в БГУИР, Минск, Беларусь. Он полностью открыт для использования[8] и хорошо изучен в контексте применения методов обработки сигналов для точного определения речевых нарушений. Однако применение данного набора для изучения когнитивных нарушений ограничено в связи с его небольшим размером и отсутствием разговорной речи.

Подводя итоги раздела, можно выделить несколько проблем речевых наборов по когнитивным нарушениям:

1) Многие речевые наборы данных являются закрытыми, либо предоставляют только отдельные поведенческие сигналы, часто только аудиосигналы, что приводит к необходимости использовать нейросетевые методы для автоматического извлечения дополнительных модальностей, например, автоматическое распознавание речи.

2) Крайне малочисленны наборы данных, которые можно использовать для автоматического определения сразу нескольких заболеваний.

3) Малое число наборов на естественных языках, отличных от английского и китайского, что ограничивает исследование кросс-языковых методов.

4) Отсутствуют единообразные протоколы сбора речевых и аудиовизуальных данных по когнитивным нарушениям, что объясняется значительно различающимися условиями записи.

5) Существующие наборы данных, за редким исключением, имеют дисбаланс в количестве экземпляров в классах обучающих данных (редки данные с когнитивными нарушениями), а также могут содержать посторонние шумы, голос интервьюера и т.д.

Все это значительно затрудняет создание систем, основанных на глубоких нейросетевых моделях и аудиовизуальных многомодальных подходах.

4. Обзор работ по многоязычному определению когнитивных нарушений в речи, представленных на соревновании TAUKADIAL. В опубликованных ранее аналитических обзорах [38, 39] были рассмотрены работы по неинвазивному

[8] https://github.com/Mak-Sim/Troparion/tree/master/SPA2019

автоматическому одно- и бимодальному определению когнитивных нарушений по аудио и тексту. Авторы данных обзоров подробно проанализировали основные признаки, методы и модели, позволившие получить наилучшие результаты на международных соревнованиях ADReSS, ADReSSo, NCMMSC's и ADReSS-M. Однако наиболее актуальное соревнование TAUKADIAL по многоязычному определению когнитивных нарушений, проводившееся в рамках международной конференции INTERSPEECH-2024, не было рассмотрено.

В ходе соревнования TAUKADIAL 2024 [27] участникам предлагалось разработать систему для автоматического кросс-языкового определения когнитивных нарушений на основе аудиоданных на английском и китайском языках. Данным на обоих языках предоставлялась метка диагноза для автоматической классификации и метка по значению шкалы MMSE для решения регрессионной задачи. Аудиозаписи не были сопровождены текстовыми транскрипциями. Лучшие результаты работы систем, представленных на соревновании, отображены в таблице 2.

Таблица 2. Сравнение результатов систем на соревновании TAUKADIAL 2024

Работа	Задача*	Модальность	Набор признаков	Модель	Значения показателей val/test
Luz S. et al. базовая работа [27]	Класс.	Аудио	wav2vec и eGeMAPS	MLP	UAR = 50,9/59,2 %
	Регр.	Текст (Google ASR)	Лингвистические	MLP	RMSE = 2,86/2,89
Duan J. et al. [41]	Класс.	Аудио и текст (Whisper large v3)	Prosody UnitY2 и XLM-RoBERTa-base	Трансформер	UAR = -/77,6 %
Favaro A. et al. [42]	Класс.	Аудио и текст (Whisper large v2 с промптом)	text2vec-base-multilingual, LaBSE, Lealla,Whisper	Ансамбль – XGC, SVC, BAGC, MLP	UAR = 70,0/75,0 %
	Регр.		Лингвистические, темпоральные, Multilingual-E5-Large, XLM-Roberta-Large-Vit-L-14	Ансамбль – XGR, SVR, BAGR, MLP	RMSE = 3,06/2,44
Hoang B. et al. [45]	Регр.	Аудио и текст (Whisper large v2 с промптом)	MFCC, BERT с аугментацией, метка языка Whisper	SVR	RMSE = 2,705/2,578
Pérez-Toro P.A. et al. [47]	Класс.	Аудио	Whisper и темпоральные	MLP	UAR = 80,0/83,3 %
	Регр.	Аудио и текст (Amazon AWS ASR)	BERT, темпоральные и Whisper	SVR	RMSE = 2,34/1,87

* Класс. – классификация КН; Регр. – Регрессия MMSE

Основным показателем точности классификации выступала невзвешенная средняя полнота (unweighted average recall, UAR). Результаты регрессии оценивались с использованием показателя корня из среднеквадратической ошибки (root mean squared error, RMSE).

В базовой системе [25], представленной на соревновании TAUKADIAL 2024, в первую очередь проверялась гипотеза возможности автоматического многоязычного или языконезависимого определения когнитивных нарушений. На рисунке 2 представлена схема универсальной архитектуры, предложенной в данной работе. Из аудиосигнала авторы выделяли 88 акустических признаков с помощью стандартизированного набора признаков eGeMAPS, включенного в программный инструментарий OpenSMILE[9]. Также авторы извлекали вектор нейросетевых представлений из аудио с помощью предобученной глубокой нейросетевой модели Wav2Vec [40]. Набор лингвистических признаков формировался на основе результатов автоматического распознавания речи и автоматической разметки частей речи и включал следующие, нормализованные по z-мере, признаки: количество лексем и слов различных частей речи, коэффициент лексического разнообразия, плотность (соотношение слов различных частей речи к суммарному числу лексем) и др. Для классификации и регрессии использовался многослойный персептрон. В результате лучшая точность классификации метки диагноза по показателю UAR составила 50,9% на валидационной выборке и 59,2% на тестовой выборке при объединении классификаторов, обученных на признаках wav2vec и eGeMAPS. Наивысшая точность предсказания значения MMSE по показателю RMSE оставила 2,86 и 2,89 на валидационной и тестовой выборках соответственно при использовании только лингвистических признаков. Другие сочетания признаков также рассматривались в исследовании, но не оказали положительного эффекта на результирующую точность классификации или регрессии.

В работе [41], в отличие от остальных методов, достигших высоких результатов, использовалось многомодальное объединение нейронных представлений посредством 6-слойной модели трансформера. Также авторы использовали автоматическое сопоставление аудио и текста посредством самообучения (self-supervised learning, SSL) для аугментации и объединения модальностей. На этапе предобработки авторы использовали автоматическое определение языка с помощью Whisper large-v3 и распознавали текст с использованием этой же модели, предварительно

[9] https://www.audeering.com/research/opensmile/

указывая промпт для сохранения речевых сбоев в записях. Для извлечения текстовых признаков авторы использовали предобученную нейросетевую модель XLM-RoBERTa-base. Для извлечения аудиопризнаков было предложено использовать модель перевода речи в латентное пространство с учетом интонации Prosody UnitY2 [10]. Модель, использующая объединение модальностей и аугментацию сопоставлением аудио и текста, достигла UAR = 77,6%, продемонстрировав тем самым один из лучших результатов на соревновании TAUKADIAL.

Рис. 2. Архитектура базовой системы определения когнитивных нарушений с помощью языконезависимых признаков по речи [27]

[10] https://github.com/facebookresearch/seamless_communication

В исследовании [42] авторы сделали значительный акцент на поиск оптимальных сочетаний интерпретируемых и неинтерпретируемых признаков. В базовый процесс предобработки были внесены следующие изменения: аудиозаписи приводились к частоте дискретизации 16 кГц и нормализовались по корню из среднеквадратической громкости (root mean squared loudness); для автоматического определения языка использовалась нейросетевая модель модель Whisper large-v2 [43], которая также использовалась для автоматического распознавания, но с добавлением промптов на английском и китайском языках для сохранения значимых для медицинского контекста речевых сбоев. Авторы исследовали следующие интерпретируемые признаки:

– признаки темпа на основе детектора речевой активности (voice activity detection, VAD) из инструментария py-webrtcvad и библиотеки анализа интонационных признаков DigiPsych Prosody;

– акустические признаки из библиотеки Simple Speech Features[11];

– текстовые признаки, извлекаемые с помощью инструментария Linguistic Feature Toolkit[12] [44].

К неинтерпретируемым рассмотренным признакам относятся:

– мел-частотные кепстральные коэффициенты (Mel-frequency Cepstral Coefficients, MFCC), включенные в набор признаков ComParE библиотеки OpenSMILE;

– нейросетевые векторы акустических моделей Wav2vec и Whisper, усредненные по временной оси для каждой записи для получения одномерного вектора;

– нейросетевые векторы больших языковых моделей (Large Language Model, LLM).

Лучшие результаты классификации были получены с применением только моделей, обученных на нейросетевых признаках текстовых моделей и Whisper, и составили 70% и 75% по показателю UAR на валидационной и тестовой выборках соответственно. Лучшие результаты при решении задачи регрессии показал ансамбль классификаторов на основе бэггинга, градиентного бустинга и метода опорных векторов, обученных на экспертных и нейросетевых признаках, составив 3,06 и 2,44 по RMSE на валидационной и тестовой выборках соответственно. Использование только интерпретируемых лингвистических и темпоральных признаков с теми же регрессорами,

[11] https://github.com/uzaymacar/simple-speech-features
[12] https://lftk.readthedocs.io/en/latest/

что и в лучшем подходе, также показало улучшение результатов по сравнению с базовыми значениями, незначительно уступающее лучшему результату на тестовой выборке – 3,06 и 2,62 RMSE для валидационной и тестовой выборок соответственно.

В исследовании [45] авторы предположили, что можно успешно применить методы аугментации текста посредством автоматического перевода для решения проблемы ограниченности размеров обучающей выборки и обеспечения универсальности модели по двум языкам. Для аугментации язык записи определялся автоматически с применением модели Whisper large v2 и автоматически распознавался этой же моделью. Для перевода авторы использовали многоязычную нейросетевую модель M2M100 [46], после чего из данных извлекались нейросетевые признаки с помощью англоязычной или многоязычной модели BERT. Также авторы извлекали акустические признаки из аудиозаписей с помощью набора признаков eGeMAPS из библиотеки OpenSMILE. Для уменьшения размерности признаков использовался метод LASSO. Для регрессии авторы использовали – регрессию методом опорных векторов и случайного леса. Полученные результаты предсказания метки RMSE превзошли базовые и составили 2,705 и 2,578 RMSE на валидационной и тестовых выборках соответственно.

В работе [47] были продемонстрированы наилучшие результаты в рамках соревнования TAUKADIAL как по задаче классификации когнитивных нарушений, так и по задаче предсказания шкалы MMSE. На этапе предобработки данных авторы обнаружили, что многие англоязычные записи могут содержать речь интервьюера, а не только респондента, из-за чего было принято решение использовать функции VAD, определения говорящего, определения языка и автоматического распознавания речи, предлагаемые коммерческим решением Amazon web services (AWS) transcribe [13]. В качестве признаков использовались нейросетевые признаки, многоязычных языковых моделей RoBERTa и BERT, а также векторы акустических моделей Whisper и Wav2Vec XLSR-53. Также авторы предположили значимость использования данных по темпу в данной задаче, поэтому в дополнение к нейросетевым признакам рассчитывались 16 темпоральных признаков на основе VAD, такие как число пауз и речевых сегментов в секунду, соотношение числа речевых сегментов и пауз, средняя длительность пауз и речевых сегментов и т.д. Для решения задачи классификации авторы рассматривали многослойный персептрон и метод опорных

[13] https://www.amazonaws.cn/en/transcribe/

векторов. Для задачи предсказания значения MMSE рассматривались линейная регрессия и метод опорных векторов. Наилучшие результаты классификации были получены при использовании совмещенного вектора темпоральных признаков и скрытого слоя Whisper для обучения многослойного персептрона, а также использования неполного набора данных для обучения, включавшего только задачи на описание изображений. Значения UAR составили 80,0% и 83,3% на валидационной и тестовой выборках соответственно. Наилучшие результаты по регрессии составили 2,34 и 1,87 по показателю RMSE и были получены при использовании совмещенного вектора BERT, темпоральных признаков и скрытого слоя Whisper для обучения регрессора методом опорных векторов на неполном наборе данных, содержащим только первую подвыборку описаний изображений.

Следует отметить, что в большинстве работ, представленных на соревнованиях, были использованы неглубокие нейросетевые классификаторы или вероятностные модели (деревья решений, метод опорных векторов, градиентный бустинг). Это обусловлено сравнительно малыми размерами исходных данных. Также отмечается интерес к поиску новых сочетаний интерпретируемых, в частности статистических характеристик темпа, и неинтерпретируемых признаков. Несмотря на то, что во многих случаях методы, основанные на нейросетевых неинтерпретируемых признаках, показывают более высокие результаты по сравнению с интерпретируемыми аналогами, отставание интерпретируемых методов по точности не настолько значительно.

5. Обзор работ по визуальному и многомодальному определению когнитивных нарушений. В этом разделе предлагается анализ исследований, выполненных вне соревнований и включающих методы, основанные на визуальной модальности, менее распространенные в данной области в силу ограниченного числа аудиовизуальных корпусов. В таблице 3 представлены основные работы, выполненные вне соревнований и включающие исследования по применению видео и многомодальных (аудио, текст, видео и др.) методов. Точность моделей в таблице приводится преимущественно по показателю точности (Accuracy) в целях обеспечения сопоставимости результатов, также в одном случае используется значение показателя площади под кривой (area under curve, AUC) для оценки модели, поскольку авторы не привели значение точности.

Таблица 3. Сравнение результатов систем, основанных на видео и многомодальных признаках

Работа	Набор данных	Модальность	Набор признаков	Модель	Точность
Zheng C. et al. 2023 [48]	PROMPT	Видео	Коды лицевых движений	SVM	0,71
			Гистограммы направленных градиентов	LSTM	0,79
Okunishi T. et al. 2025 [49]	PROMPT	Видео	Коды лицевых движений, классы эмоций, валентность возбуждение и лицевые эмбеддинги	Дерево решений	0,93 (AUC)
Mu X. et al. 2024 [50]	I-CONECT	Аудио, видео, текст (Whisper)	LLAMA-65B, RoBERTa Sentiment, WavLM, DINOv2, лицевые эмоции, коды лицевых движений, rPPG, акустические и демографические признаки	Позднее объединение детерминированных моделей голосованием большинством	0,58
		Аудио, текст (Whisper)	LLAMA-65B, RoBERTa Sentiment, WavLM, акустические и демографические признаки	Позднее объединение детерминированных моделей голосованием большинством	0,63

В работе [48] представлено одно из первых систематических исследований по автоматическому определению когнитивных нарушений с использованием видеомодальности. Авторы исследовали использование кодов лицевых движений с классификацией методом опорных векторов SVM, а также главные компоненты по лицевым полигональным сеткам и гистограммам направленных градиентов при обучении модели с долгой краткосрочной памятью (long-short term memory, LSTM) в применении к набору данных PROMPT на японском языке. Наилучший результат классификации с точностью 78% был получен при использовании гистограмм направленных компонент, хотя авторы отмечают, что данный подход может быть подвержен влиянию освещения на изображении. Использование кодов лицевых движений также продемонстрировало относительно высокую точность, составившую 71% процент. Кроме того, установлено, что удаление незначимых кодов лицевых движений не оказало

существенного влияния на результат. В целом исследование подтвердило возможность использования видеомодальности для выявления когнитивных нарушений.

В работе [49] в качестве признаков для обучения дерева решений авторы использовали коды лицевых движений и автоматически определенные классы эмоций, вычисленные с помощью библиотеки OpenFace, а также метки валентности-включенности и нейросетевые эмбеддинги. Результирующее значение AUC = 0,93. При этом использование дерева решений позволило сохранить интерпретируемость результатов относительно значимых признаков.

В работе [50] представлен многомодальный подход к определению когнитивных нарушений и предсказанию оценок психологического благополучия на материале корпуса I-CONECT. Архитектура предложенной системы представлена на рисунке 3. В исследовании используется позднее объединение модальностей [51]. Признаки различных модальностей, в том числе демографические (раса, пол, возраст), обрабатываются отдельными моделями: с помощью бинарных классификаторов, логистической регрессии и/или классификаторов с градиентным усилением для классификации шкал, после чего производится объединение на уровне меток классификаторов посредством голосования большинством. Лучшая точность определения КН составила 63% без использования данных лицевой модальности и сердечно-сосудистых признаков. При объединении всех признаков точность снизилась до 58%. В задаче предсказания значения шкалы MoCA модель также показала не самые высокие результаты, достигнув максимальной точности в 58% при использовании всех модальностей. Однако в задаче предсказания значений шкалы клинической деменции (CDR, Clinical Dementia Rating) точность модели составила 74%.

Следует отметить, что изучение видеомодальности для автоматического определения когнитивных нарушений представлено меньшим количеством работ в сравнении с аудио и текстовыми подходами. В рассмотренных нами работах точность в среднем уступает результатам, полученным на аудио и текстовом материале. Однако исследование видеомодальности представляется перспективным направлением в контексте разработки языково-независимых методов. Особый интерес представляют предварительные исследования по извлечению промежуточных нейросетевых признаков, таких как эмоциональное состояние и автоматически распознанные физиологические признаки на основе удаленной фотоплетизмографии и иные.

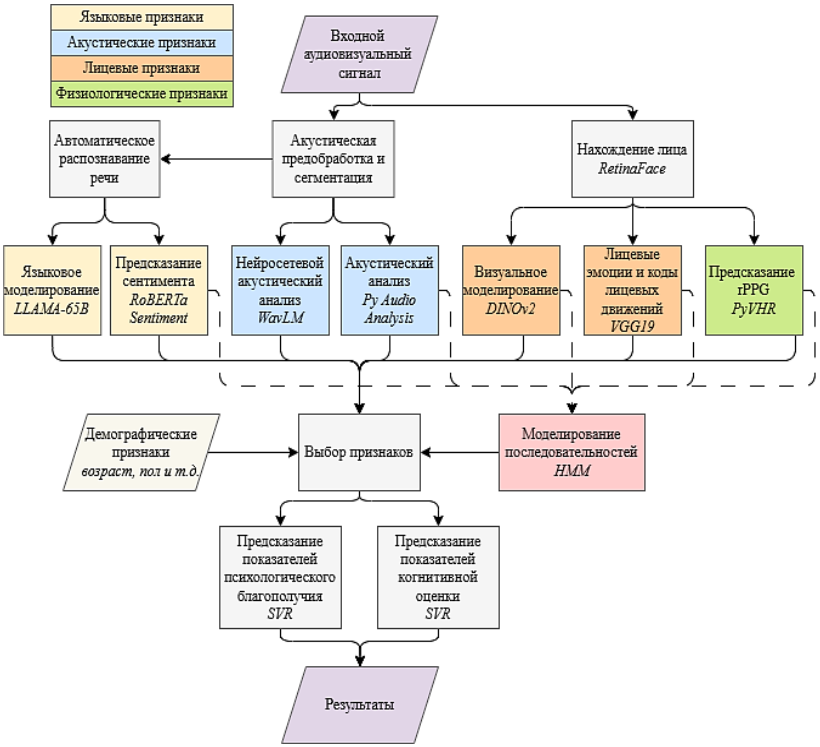

Рис. 3. Архитектура многомодальной системы предсказания показателей когнитивной оценки и психологического благополучия [50] (на рисунке курсивом указаны признаки, инструментарии и модели, использовавшиеся в реализации отдельных модулей системы)

6. Применение языковых макродескрипторов для определения когнитивных нарушений. Как упоминалось в разделе 3, некоторые из симптомов заболеваний, хотя они и естественно отмечаются специалистами и в некоторых случаях даже не специалистами, сравнительно трудно отразить с помощью интерпретируемых количественных лингвистических или акустических признаков. Для того, чтобы решить данную проблему, некоторые авторы последних работ предлагают использовать большие языковые модели для извлечения промежуточных высокоуровневых признаков (макродескрипторов), формализация которых затруднительна.

Впервые данный метод был использован в работе [52], где предлагалось использовать LLM для оценки беглости речи людей.

Предложенная в работе архитектура системы представлена на рисунке 4. Полученные метки беглости речи использовались для обучения нейросетевой модели, вместе с векторными нейросетевыми представлениями аудио и текста, что лишило результирующую систему возможности интерпретации полученных результатов. Наилучшее значение по показателю F-меры составило 87,25% при кросс-валидации на материале набора ADReSSo.

Авторы работы [53] предложили развитие идеи, сравнив эффективность использования больших языковых моделей в качестве прямой системы, определяющей наличие болезни Альцгеймера у пациента, с использованием данных моделей для выделения промежуточных высокоуровневых текстовых признаков и последующего обучения интерпретируемых стохастических моделей классификации. Исследование проводилось на наборе данных ADReSS, содержащем текстовые расшифровки, но рассматривался также вариант, при котором текстовые расшифровки отсутствовали, в результате чего применялись различные системы автоматического распознавания речи Whisper-large, Wav2Vec-large или иные. Наименьшую ошибку распознавания слов в речи WER = 26,9% продемонстрировал Whisper, хотя он склонен опускать заполненные паузы-хезитации, важные для определения заболеваний. Рассматривались такие LLM, как Mistral 7B[14], Mixtral 8x7B[15] и GPT-3.5[16]. Наилучшие и наиболее стабильные результаты были достигнуты при использовании LLM для извлечения промежуточных признаков и предсказания болезни с предоставлением модели информации о работе в медицинской области. Значение F1-меры = 79,2% при обучении классификатора на основе случайного леса, и 81,3% при обучении классификатора на основе опорных векторов с использованием макродескрипторов и предсказания заболевания модели Mistral 7B. Интересно отметить, что языковые модели большего размера демонстрировали в среднем более низкие и менее стабильные результаты.

[14] https://mistral.ai/news/announcing-mistral-7b
[15] https://huggingface.co/blog/mixtral
[16] https://openai.com/index/introducing-chatgpt-and-whisper-apis/

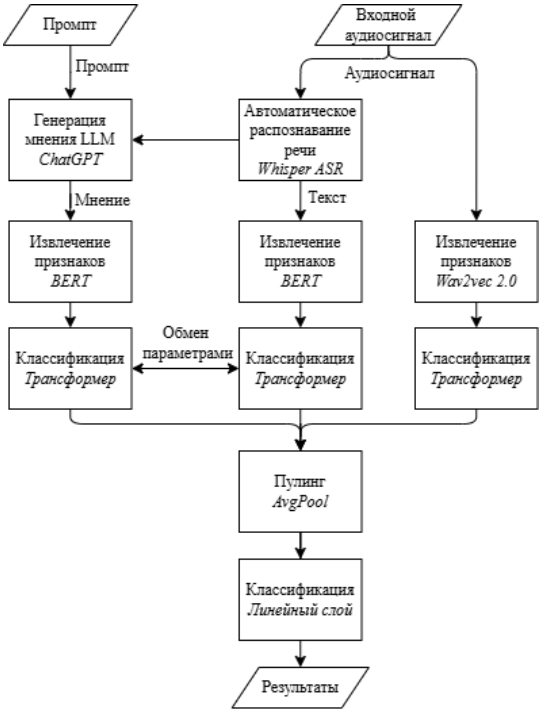

Рис. 4. Архитектура многомодальной системы предсказания показателей когнитивной оценки с применением LLM [52] (на рисунке курсивом указаны признаки, инструментарии и модели, использовавшиеся в реализации отдельных модулей системы)

В работе [14] авторы предложили многомодальный и интерпретируемый подход для определения нескольких заболеваний: слабых когнитивных нарушений, деменции при БА и БП. Эксперименты проводились на следующих наборах данных: Pitt по данным людей с когнитивными нарушениями на английском, PC GITA по данным людей с болезнью Паркинсона на испанском и немедицинском наборе CLAC [54] по данным оценки владения английским для расширения контрольной группы здоровых носителей. Текстовая транскрипция из речи извлекалась автоматически с помощью Whisper-large. При этом использовались различные интерпретируемые акустические, связанные с качеством голоса и формой речевого тракта, и темпоральные признаки, а также лингвистические количественные, например пропорции различных частей речи в тексте, плотность содержания и другие из библиотеки

BlaBla [17] и, основанные на нейросетевых моделях, промежуточные признаки, например связность текста на основе косинусной близости векторов нейросетевой модели All-Mpnet-Base-v2[18], сентимент-анализ моделью TextBlob [19] и анализ цепочек кореферентности, последовательностей слов в тексте, которые относятся к одному и тому же обозначаемому, на основе нейросетевых языковых моделей из инструментария wl-coref[20] [55].

Для снижения размерности вектора признаков в [14] использовали корреляционный анализ, то есть удалялись сильно коррелированные признаки, если они оказывались незначимы для речи в выборке людей с когнитивными нарушениями и в выборке без когнитивных нарушений. Признаки группировались в кластеры с помощью иерархической кластеризации, основанной на корреляции Пирсона между всеми парами признаков. Для повышения устойчивости к шумам в разных наборах данных, эталонные признаки выбирались на основе сходства стандартного отклонения их распределений в выборках по здоровым людям и пациентам. Признак с наиболее схожим стандартным отклонением в обеих группах в каждом кластере был определен как эталонный. Средние значения не учитывались, так как их можно скорректировать путем добавления константы.

В качестве классификатора используется нейронная аддитивная модель (Neural Additive Model, NAM) [56], идея которой заключается в том, чтобы применить линейную аппроксимацию функций, представляемых в виде нейронных сетей с одним входом, соответствующих одному признаку, а затем обучить нейронные сети совместно с обратным распространения ошибки. В результате для каждого предсказания можно восстановить вклад входных признаков и связать наиболее значимые акустические или лингвистические признаки с соответствующими симптомами нарушений, понятными специалисту. Точность предложенной модели составила 84,3% на тестовой выборке набора по БА и 75% на тестовой выборке набора по БП. В обоих случаях признаки для обучения нормализовались с помощью нормализации с помощью размаха (MinMax). При этом предложенное решение сохраняет интерпретируемость и демонстрирует сопоставимые результаты с наиболее современными

[17] https://github.com/novoic/blabla
[18] https://huggingface.co/sentence-transformers/all-mpnet-base-v2
[19] https://textblob.readthedocs.io/en/dev/
[20] https://github.com/vdobrovolskii/wl-coref

решениями, основанными на нейросетевых признаках для данной задачи.

Необходимо также отметить, что существуют несколько эталонных наборов данных для оценки применимости LLM относительно широкого спектра медицинских задач, в частности автоматических ответов на вопросы и проверки полноты данного ответа, в том числе и на вопросы, связанные с когнитивными нарушениями. Несколько наиболее популярных наборов для оценки возможностей генеративных моделей по решению медицинских задач:

- MedQA [57], MedMCQA [58] – англоязычные наборы данных только по ответам на вопросы (question-answering, QA);
- HealthBench [59] – англоязычный набор данных с различными типами задач: QA, генерация, классификация;
- RuMedBench [60] – русскоязычный набор данных с различными типами задач: QA, генерация, классификация.

В результате появления доступных материалов для обучения начали возникать и специализированные медицинские LLM. Далее рассмотрим несколько примеров подобных моделей.

Одна из первых работ, предлагающая создание многозадачной и многомодальной медицинской LLM, – Med-PaLM M [61] на основе собственного набора MultiMedBench. Результаты предложенной системы значительно превзошли результаты неспециализированной PaLM-E [62] во всех задачах (ответы на вопросы, автоматическое реферирование, классификация изображений и др.), хотя показатели специализированных систем эта модель не превзошла в задачах на основе текстовой модальности, либо превзошла только незначительно в задачах на основе визуальной модальности. Также авторы отмечают, что модель плохо справляется с адаптацией на основе небольшого количества примеров (few-shot learning, FSL).

В работе [63] предлагается создание серии специализированных медицинских нейросетевых языковых моделей с открытым доступом и количеством обучаемых параметров менее 10 миллиардов: Meerkat 7B и 8B. Результаты продемонстрировали, что относительно небольшое количество параметров и тщательный подбор данных для обучения цепочкам рассуждений на основе медицинских учебников позволяют значительно превзойти результаты многих специализированных и неспециализированных больших языковых моделей на медицинских эталонных данных. Однако, представленные модели все еще значительно отстают по количественным показателям от коммерческих LLM, таких как GPT-4o[21] и o1[22].

[21] https://openai.com/index/hello-gpt-4o/

В перспективе, применение специализированных медицинских LLM могло бы показать лучшие результаты по сравнению с неспециализированными моделями, ранее исследованными в задаче по определению когнитивных нарушений.

Применение идеи макродескрипторов на основе глубоких нейронных сетей представляется одной из наиболее интересных в контексте разработки технологий интерпретируемого искусственного интеллекта для здравоохранения. При этом предложенные решения на данный момент не рассматривали макродескрипторы для анализа многоязычных данных и трудноизмеримых качественных характеристик аудио и видео, например, для анализа эмоционального состояния, что оставляет пространство для будущих исследований.

7. Систематизация информативных признаков и классификаторов, используемых при разработке многомодальных систем автоматического определения когнитивных нарушений. В ходе проведенного обзора был выявлено несколько основных тенденций в выборе признаков и моделей для классификации.

Для учета дисбаланса в работах часто используются показатели точности моделей, устойчивые к дисбалансу. Возможно также применение методик аугментации данных и кросс-корпусных методов. Также важен выбор наиболее информативных признаков. В рассмотренных работах упоминались следующие: метод главных компонент, лассо-регрессия и др.

Для извлечения признаков использовались методы, которые можно классифицировать по модальностям, на основе которых они извлекаются: аудио, видео и текст, среди которых можно выделить разделение на нейросетевые методы и программные инструментарии, позволяющие извлекать экспертные или нейросетевые признаки, они представлены на рисунке 5.

Несмотря на то, что изменения в поведении людей с когнитивными нарушениями могут проявляться в аудио в форме отклонений тона голоса, артикуляции, беглости речи и др. и в видео в форме отклонений в проявлении эмоции и снижения мимической активности, наиболее чувствительной к нарушениям во многих работах оказывались признаки текстовой модальности. С обработкой теста во многих наборах тесно связана задача автоматического распознавания речи и речи с нарушениями, поскольку в наборах ручная транскрипция может отсутствовать. Лучшие модели распознавания речи показывают значительное снижение точности работы на данных с когнитивными нарушениями. В исследовании [64],

[22] https://openai.com/index/introducing-openai-o1-preview/

рассматривались значения WER для предобученных моделей на основе нейросетевых архитектур Wav2Vec и Whisper, полученные на наборе ADReSS. Whisper-large-v3 продемонстрировал WER=40,07% на выборке речи людей с когнитивными нарушениями и 28,47% на контрольной группе, тогда как wav2vec2-large-960h-lv60-self[23] показал 50,48% и 35,97% на речи людей с когнитивными нарушениями и контрольной группе соответственно. В исследовании [66] отмечается заметное влияние точности распознавания речи на автоматическое определение когнитивных нарушений, поэтому необходимо рассматривать также задачу дообучения моделей на медицинских данных.

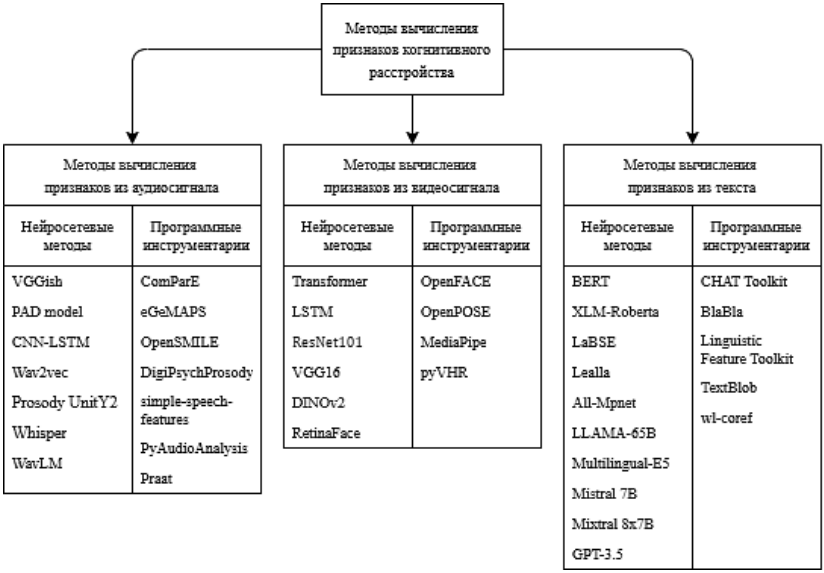

Рис. 5. Систематизация методов для извлечения информативных признаков из многомодальных данных

Среди рассмотренных работ можно выделить регрессионные и классификационные системы, в которых используются как детерминированные машинные классификаторы, так и нейросетевые классификаторы. На рисунке 6 представлены основные, использовавшиеся в работах классификаторы. Важно отметить, что традиционные детерминированные классификаторы не теряют

[23] https://huggingface.co/facebook/wav2vec2-large-960h-lv60-self

популярности, несмотря на активное развитие нейросетевых методов. Данная тенденция связана с малыми размерами данных для обучения и ориентацией на использование интерпретируемых моделей в медицинских задачах. Также можно заметить, что в работах совсем недавно стали изучаться LLM для извлечения интерпретируемых признаков.

Рис. 6. Систематизация методов машинной классификации, используемых при определении когнитивных нарушений

Важно отметить, что исследование применения современных интерпретируемых подходов, таких как: нейронные аддитивные модели; модели анализа выживаемости, подходящие для предсказания когнитивного снижения через время [66]; интерпретируемые графовые нейронные модели [67] – находится на начальном этапе.

Находят свое применение инструменты интерпретации предсказаний трудноинтерпретируемых моделей. Например, в работе [68], метод SHAP применяется для анализа влияния признаков на результат предсказания для одного человека, полученный с помощью модели XGBoost [24]. Данный метод позволяет получить индивидуальную оценку риска заболевания, что важно при диагностике конкретного пациента.

[24] https://xgboost.readthedocs.io/en/latest/

8. Заключение. В статье представлен аналитический обзор современных научных исследований, посвященных разработке автоматических систем определения когнитивных нарушений людей.

В результате выполненного аналитического обзора можно сделать несколько основных выводов:

1) Объединение текстовой модальности с невербальными акустическими, например, учетом пауз, темпа, или лицевыми признаками, например, эмоциональными признаками, позволяет улучшить результаты, но в силу отсутствия текста во многих наборах требуется учитывать качество моделей распознавания речи.

2) Анализ лексической составляющей позволяет обнаружить различные симптомы, например, снижение лексического разнообразия, трудности в нахождении слов и др.

3) Конкурентоспособные решения представляются, в том числе, с применением экспертных признаков и линейных статистических моделей, что потенциально может обеспечить интерпретируемость полученных результатов.

На основе проведенного анализа можно сформулировать потенциальные требования, которые могут быть выдвинуты к разрабатываемым автоматическим системам определения когнитивных нарушений:

1) Необходимо использовать средства нормализации и борьбы со статистическими выбросами и шумами. Большинство речевых и аудиовизуальных наборов по когнитивным нарушениям отличаются малыми объемами, несбалансированностью, неоднообразными условиями записи и высокой внутридикторской вариативностью, из-за чего данные могут быть зашумлены и модели могут переобучаться на отдельных наборах данных.

2) Необходимо отдать предпочтение интерпретируемым моделям классификации, экспертным наборам признаков, а также инструментам интерпретации (например, SHAP) нейросетевых моделей: глубоких нейронных сетей, больших языковых моделей и др. Применение моделей типа «черный ящик» может потенциально привести к ложным пропускам и срабатываниям, т.е. отсутствию своевременного медицинского вмешательства или наоборот к вмешательству в здоровом случае. Интерпретация используемых моделей позволяет специалисту отсечь незначимые случайные закономерности, которые могли возникнуть из-за ограничений обучающих данных.

3) Рекомендуется использовать и интегрировать максимально возможное количество модальностей. Отдельная модальность (текст,

аудио, видео и иные) зачастую не отражают всех особенностей поведения человека, что может привести автоматические средства к ошибкам первого или второго рода (ложным срабатываниям или пропуску нарушений).

В дальнейших исследованиях мы планируем разработать собственный прототип автоматической системы определения когнитивных нарушений по речи, учитывающий разработанные требования, и исследовать на практике акустические, лексические и визуальные лицевые признаки когнитивных нарушений, выявленные в ходе данного обзора и методы их аугментации.

Мы хотели бы посвятить эту работу памяти Рафаэля Мидхатовича Юсупова, который всегда активно поддерживал направления исследований и разработок лаборатории речевых и многомодальных интерфейсов СПИИРАН (СПб ФИЦ РАН). Рафаэль Мидхатович также являлся одним из основоположников становления и развития в России новой научной области – интеллектуальных многомодальных человеко-машинных интерфейсов и систем [69]. Светлая ему память.

Литература

1. Парфенов В.А. Ведение пациентов с когнитивными нарушениями // Неврология, нейропсихиатрия, психосоматика. 2023. Т. 15(1). С. 97–102.
2. Jonell P. et al. Multimodal Capture of Patient Behaviour for Improved Detection of Early Dementia: Clinical Feasibility and Preliminary Results // Frontiers in Computer. Sciences. 2021. vol. 3. DOI: 10.3389/fcomp.2021.642633.
3. Javeed A., Dallora A.L., Berglund J.S., et al. Machine Learning for Dementia Prediction: A Systematic Review and Future Research Directions. Journal of Medical Systems. 2023. vol. 47(17). DOI: 10.1007/s10916-023-01906-7.
4. Ляксо Е.Е., Фролова О.В., Гречаный С.В., Матвеев Ю.Н., Верхоляк О.В., Карпов А.А. Голосовой портрет ребенка с типичным и атипичным развитием // СПб.: Издательско-полиграфическая ассоциация высших учебных заведений. 2020. 204 с.
5. Величко А.Н., Карпов А.А. Аналитический обзор систем автоматического определения депрессии по речи // Информатика и автоматизация. 2021. № 20(3). С. 497–529. DOI: 10.15622/ia.2021.3.1.
6. Боголепова А.Н. и др. Клинические рекомендации «Когнитивные расстройства у пациентов пожилого и старческого возраста» // Журнал неврологии и психиатрии им. С.С. Корсакова. 2022. Т. 121(10-3). С. 6–137.
7. Ganguli M., Chang C.C., Snitz B.E., Saxton J.A., Vanderbilt J., Lee C.W. Prevalence of mild cognitive impairment by multiple classifications: The Monongahela-Youghiogheny Healthy Aging Team (MYHAT) project // American Journal of Geriatric Psychiatry. 2010. vol. 18(8). pp. 674–683. DOI: 10.1097/JGP.0b013e3181cdee4f.
8. Larrieu S., Letenneur L, Orgogozo J.M., Fabrigoule C., Amieva H., Le Carret N., Barberger-Gateau P., Dartigues J.F. Incidence and outcome of mild cognitive impairment in a population-based prospective cohort // Neurology. 2002. vol. 59(10). pp. 1594–1599. DOI: 10.1212/01.wnl.0000034176.07159.f8.

9. Roberts R.O, Geda Y.E., Knopman D.S., Cha R.H., Pankratz V.S., Boeve B.F., Ivnik R.J., Tangalos E.G., Petersen R.C., Rocca W.A. The Mayo Clinic Study of Aging: design and sampling, participation, baseline measures and sample characteristics // Neuroepidemiology. 2008. vol. 30(1). pp. 58–69. DOI: 10.1159/000115751.
10. Чердак М.А. и др. Распространенность когнитивных расстройств у пациентов старшего возраста в Российской Федерации // Журнал неврологии и психиатрии им. С.С. Корсакова. Спецвыпуски. 2024. Т. 124(4-2). С. 5–11.
11. Wallensten J., Ljunggren G., Nager A., Wachtler C., Bogdanovic N., Petrovic P., Carlsson A.C. Stress, depression, and risk of dementia – a cohort study in the total population between 18 and 65 years old in Region Stockholm // Alzheimer's Research & Therapy. 2023. vol. 15(161). DOI: 10.1186/s13195-023-01308-4.
12. Локшина А.Б., Гришина Д.А. Терапия некогнитивных нервно-психических расстройств при болезни Альцгеймера // Неврология, нейропсихиатрия, психосоматика. 2021. Т. 13(6). С. 132–138. DOI: 10.14412/2074-2711-2021-6-132-138.
13. Nestor P.J., Scheltens P., Hodges J.R. Advances in the early detection of Alzheimer's disease // Nature Medicine. 2004. vol. 10. pp. S34–S41. DOI: 10.1038/nrn1433.
14. Botelho C., Abad A., Schultz T., Trancoso I. Speech as a biomarker for disease detection // IEEE Access. 2024. vol. 12. pp. 184487–184508.
15. Yamada Y., Shinkawa K., Ishikawa T., Nishimura M., Nemoto M., Tsukada E., Ota M., Nemoto K., Arai T. Multimodal behavioral analysis for early detection of Alzheimer's disease: A preliminary result: Neuropsychiatry and behavioral neurology/Assessment/Measurement of neuropsychiatric/Behavioral and psychological symptoms // Alzheimer's & Dementia. 2020. vol. 16. DOI: 10.1002/alz.042897.
16. Boschi V., Catricala E., Consonni M., Chesi C., Moro A., Cappa S.F. Connected speech in neurodegenerative language disorders: a review // Frontiers in psychology. 2017. vol. 8. DOI: 10.3389/fpsyg.2017.00269.
17. Hecker P., Steckhan N., Eyben F., Schuller B.W., Arnrich B. Voice analysis for neurological disorder recognition–a systematic review and perspective on emerging trends // Frontiers in Digital Health. 2022. vol. 4. DOI: 10.3389/fdgth.2022.842301.
18. Voleti R., Liss J.M., Berisha V. A review of automated speech and language features for assessment of cognitive and thought disorders // IEEE journal of selected topics in signal processing. 2019. vol. 14(2). pp. 282–298. DOI: 10.1109/jstsp.2019.2952087.
19. Byers A.L., Yaffe K. Depression and risk of developing dementia // Nature Reviews Neurology. 2011. vol. 7(6). pp. 323–331. DOI: 10.1038/nrneurol.2011.60.
20. Braun F., Bayerl S.P., Perez-Toro P.A., Honig F., Lehfeld H., Hillemacher T., Noth E., Bocklet T., Riedhammer K. Classifying Dementia in the Presence of Depression: A Cross-Corpus Study // Proceedings of Interspeech 2023. vol. 2023. pp. 2308–2312. DOI: 10.21437/Interspeech.2023-1997.
21. Jonell P., et al. Multimodal Capture of Patient Behaviour for Improved Detection of Early Dementia: Clinical Feasibility and Preliminary Results // Frontiers in Computer Science. 2021. vol. 3. DOI: 10.3389/fcomp.2021.642633.
22. Lanzi A.M., Saylor A.K., Fromm D., Liu H., MacWhinney B., Cohen M.L. DementiaBank: Theoretical Rationale, Protocol, and Illustrative Analyses // American Journal of Speech-Language Pathology. 2023. vol. 32(2). pp. 426–438. DOI: 10.1044/2022_AJSLP-22-00281.
23. Becker J.T., Boller F., Lopez O.L., Saxton J., McGonigle K.L. The natural history of Alzheimer's disease: description of study cohort and accuracy of diagnosis // Archives of Neurology. 1994. vol. 51(6). pp. 585–594. DOI: 10.1001/archneur.1994.00540180063015.
24. Luz S., Haider F., Fuente S.d.l., Fromm D., MacWhinney B. Alzheimer's Dementia Recognition through Spontaneous Speech: The ADReSS Challenge // Proceedings Interspeech. 2020. pp. 2172–2176. DOI: 10.21437/Interspeech.2020-2571.

25. MacWhinney B. The CHILDES Project: Tools for Analyzing Talk // Child Language Teaching and Therapy. 1992. vol. 8(2). pp. 217–218. DOI: 10.1177/026565909200800211.
26. Luz S., et al. Alzheimer's dementia recognition through spontaneous speech // Frontiers in Computer Science. 2021. vol. 3.
27. Luz S., et al. Connected Speech-Based Cognitive Assessment in Chinese and English // Proceedings Interspeech. 2024. pp. 947–951. DOI: 10.21437/Interspeech.2024-1807.
28. Chen X., Zhang W., Ma Y. Raw Waveform-Based End-to-End Alzheimer's Disease Detection Method // Acta Electronica Sinica. 2023. vol. 51. no. 12. pp. 3582–3590.
29. Псарёва Н.Н. Беглость спонтанной речи как предиктор мягкого когнитивного снижения // ВКР по программе бакалавриата. М.: ВШЭ, 2024.
30. Karakostas A., Briassouli A., Avgerinakis K., Kompatsiaris I., Tsolaki M. The dem@care experiments and datasets: a technical report // arXiv preprint arXiv:1701.01142. 2016.
31. Kishimoto T. et al. PROMPT collaborators. The project for objective measures using computational psychiatry technology (PROMPT): Rationale, design, and methodology // Contemporary Clinical Trials Communications. 2020. vol. 19. DOI: 10.1016/j.conctc.2020.100649.
32. Poor F.F. et al. Prediction of Mild Cognitive Impairment Using a Hybrid Audio-Visual Approach: An I_CONECT Study // Alzheimer's & Dementia Journal. 2023. vol. 19. DOI: 10.1002/alz.074808.
33. Correia J., Teixeira F., Botelho C., Trancoso I., Raj B. The in-the-Wild Speech Medical Corpus // 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2021. pp. 6973–6977. DOI: 10.1109/ICASSP39728.2021.9414230.
34. Orozco J.R. et al. New Spanish speech corpus database for the analysis of people suffering from Parkinson's disease // International Conference on Language Resources and Evaluation (LREC). 2014. pp. 342–347.
35. Gratch J. et al. The Distress Analysis Interview Corpus of human and computer interviews // International Conference on Language Resources and Evaluation (LREC). 2014. pp. 3123–3128.
36. DeVault D. et al. SimSensei kiosk: A virtual human interviewer for healthcare decision support // Proceedings of the 13th International Conference on Autonomous Agents and Multiagent Systems (AAMAS'14). 2014. pp. 1061–1068.
37. Vashkevich M., Rushkevich Yu., Petrovsky A. Bulbar ALS Detection Based on Analysis of Voice Perturbation and Vibrato // Proceedings of inter. conf. Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA). 2019. pp. 267–272. DOI: 10.23919/SPA.2019.8936657.
38. Qi X., Zhou Q., Dong J., Bao W. Noninvasive automatic detection of Alzheimer's disease from spontaneous speech: a review // Frontiers in Aging Neuroscience. 2023. vol. 15.
39. Ding K., Chetty M., Noori Hoshyar A., Bhattacharya T., Klein B. Speech based detection of Alzheimer's disease: a survey of AI techniques, datasets and challenges // Artificial Intelligence Review. 2024. vol. 57. no. 12.
40. Babu A. et al. XLS-R: Self-supervised cross-lingual speech representation learning at scale // Proceedings of Interspeech. 2022. pp. 2278–2282. DOI: 10.21437/Interspeech.2022-143.
41. Duan J., Wei F., Li H.D., Liu J. Pre-trained Feature Fusion and Matching for Mild Cognitive Impairment Detection // Proceedings of Interspeech. 2024. pp. 962–966.
42. Favaro A., Cao T., Dehak N., Moro-Velazquez L. Leveraging Universal Speech Representations for Detecting and Assessing the Severity of Mild Cognitive Impairment Across Languages // Proceedings of Interspeech. 2024. pp. 972–976. DOI: 10.21437/Interspeech.2024-2030.

43. Radford A., Kim J.W., Xu T., Brockman G., McLeavey C., Sutskever I. Robust speech recognition via large-scale weak supervision // International conference on machine learning. PMLR 2023. 2023. pp. 28492–28518.
44. Lee B.W., Lee J. LFTK: Handcrafted features in computational linguistics // Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023). 2023. pp. 1–19. DOI: 10.18653/v1/2023.bea-1.1.
45. Hoang B., Pang Y., Dodge H., Zhou J. Translingual Language Markers for Cognitive Assessment from Spontaneous Speech // Proceedings of Interspeech. 2024. 977–981. DOI: 10.21437/Interspeech.2024-1422.
46. Fan A. et al. Beyond English-centric multilingual machine translation // Journal of Machine Learning Research. 2021. vol. 22. no. 107. pp. 1–48.
47. Pérez-Toro P.A., Arias-Vergara T., Klumpp P., Weise T., Schuster M., Noeth E., Orozco-Arroyave J.R., Maier A.. Multilingual Speech and Language Analysis for the Assessment of Mild Cognitive Impairment: Outcomes from the Taukadial Challenge // Proceedings of Interspeech. 2024. pp. 982–986. DOI: 10.21437/Interspeech.2024-2115.
48. Zheng C. et al. Detecting Dementia from Face-Related Features with Automated Computational Methods // Bioengineering (Basel). 2023. vol. 10(7). DOI: 10.3390/bioengineering10070862.
49. Okunishi T., Zheng C., Bouazizi M., Ohtsuki T., Kitazawa M., Horigome T., Kishimoto T. Dementia and MCI Detection Based on Comprehensive Facial Expression Analysis from Videos During Conversation // IEEE Journal of Biomedical and Health Informatics. 2025. vol. 29(5). pp. 3537–3548. DOI: 10.1109/JBHI.2025.3526553.
50. Mu X. et al. Detecting Cognitive Impairment and Psychological Well-being among Older Adults Using Facial, Acoustic, Linguistic, and Cardiovascular Patterns Derived from Remote Conversations // arXiv preprint arXiv:2412.14194. 2024.
51. Jiang Z., Seyedi S., Griner E., Abbasi A., Rad A.B., Kwon H., Cotes R.O., Clifford G.D. Multimodal Mental Health Digital Biomarker Analysis from Remote Interviews Using Facial, Vocal, Linguistic, and Cardiovascular Patterns // IEEE Journal of Biomedical and Health Informatics. 2024. vol. 28(3). pp. 1680–1691. DOI: 10.1109/JBHI.2024.3352075.
52. Bang J.-U., Han S.-H., Kang B.-O. Alzheimer's Disease recognition from spontaneous speech using large language models // ETRI Journal. 2024. vol. 46. no. 1. pp. 96–105. DOI: 10.4218/etrij.2023-0356.
53. Botelho C. et al. Macro-descriptors for Alzheimer's disease detection using large language models. Proc. Interspeech 2024. 2024. pp. 1975–1979.
54. Haulcy R., Glass J. CLAC: A Speech Corpus of Healthy English Speakers // Proceedings of Interspeech. 2021. pp. 2966–2970. DOI: 10.21437/Interspeech.2021-1810.
55. Dobrovolskii V. Word-Level Coreference Resolution // Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 2021. pp. 7670–7675. DOI: 10.18653/v1/2021.emnlp-main.605.
56. Agarwal R., Melnick L., Frosst N., Zhang X., Lengerich B., Caruana R., Hinton G.E. Neural additive models: Interpretable machine learning with neural nets // Advances in Neural Information Processing Systems. 2021. vol. 34. pp. 4699–4711.
57. Jin D. et al. What Disease Does This Patient Have? A Large-Scale Open Domain Question Answering Dataset from Medical Exams // Applied Sciences. 2021. vol. 11(14). DOI: 10.3390/app11146421.
58. Pal A. et al. MedMCQA: A Large-scale Multi-Subject Multi-Choice Dataset for Medical domain Question Answering // Proceedings of Machine Learning Research. 2022. vol. 174. pp. 248–260.
59. Arora R.K. et al. HealthBench: Evaluating Large Language Models Towards Improved Human Health // arXiv preprint arXiv:2505.08775. 2025.

60. Blinov P., Reshetnikova A., Nesterov A., Zubkova G., Kokh V. RuMedBench: A Russian Medical Language Understanding Benchmark // Conference on Artificial Intelligence in Medicine. 2022. pp. 383–392.
61. Tu T. et al. Towards generalist biomedical AI // NEJM Ai. 2024. vol. 1(3).
62. Driess D. et al. PaLM-E: an embodied multimodal language model. Proceedings of the 40th International Conference on Machine Learning (ICML'23). 2023. vol. 202. pp. 8469–8488.
63. Kim H. et al. Small language models learn enhanced reasoning skills from medical textbooks // NPJ Digital Medicine. 2025. vol. 8(1). DOI: 10.1038/s41746-025-01653-8.
64. Wang Y. et al. Exploiting prompt learning with pre-trained language models for Alzheimer's Disease detection // IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2023. pp. 1–5. DOI: 10.1109/ICASSP49357.2023.10095993.
65. Balagopalan A., Shkaruta K., Novikova J. Impact of ASR on Alzheimer's Disease Detection: All Errors are Equal, but Deletions are More Equal than Others // Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020). 2020. pp. 159–164. DOI: 10.18653/v1/2020.wnut-1.21.
66. Utkin L.V., Konstantinov A.V., Eremenko D.Yu., Zaborovsky V.S., Muliukha V.A. Interpretation methods for machine learning models in the framework of survival analysis with censored data: a brief over view // Computing, Telecommunications and Control. 2024. vol. 17. no. 3. pp. 22–31. DOI: 10.18721/JCSTCS.17302.
67. Alkhatib A. et al. Interpretable Graph Neural Networks for Tabular Data // Proceedings of the 27th European Conference on Artificial Intelligence (ECAI 2024). 2024. vol. 392. pp. 1848–1855.
68. Ren H., Zheng Y., Li C., Jing F., Wang Q., Luo Z., Li D., Liang D., Tang W., Liu L., Cheng W. Using Machine Learning to Predict Cognitive Decline in Older Adults From the Chinese Longitudinal Healthy Longevity Survey: Model Development and Validation Study // JMIR Aging. 2025. vol. 8. DOI: 10.2196/67437.
69. Karpov A.A., Yusupov R.M. Multimodal Interfaces of Human-Computer Interaction. Herald of the Russian Academy of Sciences. 2018. vol. 88(1). pp. 67–74. DOI: 10.1134/S1019331618010094.

Долгушин Михаил Дмитриевич — младший научный сотрудник, аспирант, лаборатория речевых и многомодальных интерфейсов, Федеральное государственное бюджетное учреждение науки «Санкт-Петербургский Федеральный исследовательский центр Российской академии наук» (СПб ФИЦ РАН). Область научных интересов: речевые технологии, автоматическое распознавание речи, интерпретируемый искусственный интеллект, компьютерная лингвистика. Число научных публикаций — 12. dolgushin.m@iias.spb.su; 14-я линия В.О., 39, 199178, Санкт-Петербург, Россия; р.т.: +7(812)328-0421.

Карпов Алексей Анатольевич — д-р техн. наук, профессор, руководитель лаборатории, лаборатория речевых и многомодальных интерфейсов, Федеральное государственное бюджетное учреждение науки «Санкт-Петербургский Федеральный исследовательский центр Российской академии наук» (СПб ФИЦ РАН). Область научных интересов: речевые технологии, автоматическое распознавание речи, обработка аудиовизуальной речи, многомодальные человеко-машинные интерфейсы, компьютерная паралингвистика. Число научных публикаций — 450+. karpov@iias.spb.su; 14-я линия В.О., 39, 199178, Санкт-Петербург, Россия; р.т.: +7(812)328-0421.

Поддержка исследований. Данное исследование выполнено при финансовой поддержке Российского научного фонда, проект № 25-11-00319, кроме раздела № 6, выполненного в рамках бюджетной темы № FFZF-2025-0003.

DOI 10.15622/ia.24.6.6

M. DOLGUSHIN, A. KARPOV

ANALYTICAL REVIEW OF SPEECH AND MULTIMODAL METHODS FOR COGNITIVE IMPAIRMENTS RECOGNITION IN PEOPLE

Dolgushin M., Karpov A. **Analytical Review of Speech and Multimodal Methods for Cognitive Impairments Recognition in People.**

Abstract. Over the past decade, there has been a noticeable increase in the number of scientific, technical, and medical publications dedicated to the automatic detection of cognitive impairments in humans based on speech and visual data. These impairments are often associated with neurodegenerative diseases such as dementia, Alzheimer's disease, Parkinson's disease, and other disorders. Despite the high prevalence of these conditions and their significant contribution to mortality and early disability, effective treatment options remain unavailable or severely limited in current medical practice. Consequently, early diagnosis and symptom alleviation have become areas of considerable research interest. Current studies focus on the development of automated and automatic systems based on quantitative and objective methods, neural network approaches, the integration of various modalities, and explainable artificial intelligence techniques. This paper presents a comprehensive review and analysis of key studies published since 2022 that address the automatic detection of cognitive impairments using unimodal and multimodal approaches. The review includes the most commonly used multimodal datasets in this domain, such as ADReSS, ADReSSo, and TAUKADIAL. It discusses state-of-the-art methods for detecting cognitive impairments from various modalities, including those presented in international competitions such as TAUKADIAL-2024, as well as methods developed outside of such events. According to competition results, the most effective approaches to recognizing cognitive impairments are ensemble probabilistic models trained on explainable hand-crafted features and neural features extracted from text and audio data. The review also explores multimodal approaches that incorporate visual modalities for training deep neural networks. A new direction in the field is examined, namely, the applicability of large language models to the analysis of medical texts and interpretable disease prediction. The paper systematizes methods for extracting informative features and the classifiers employed. Based on the review, key requirements for systems aimed at the automated detection of cognitive impairments are formulated.

Keywords: automatic detection of cognitive impairment, speech technologies in healthcare, explainable artificial intelligence, machine learning.

References

1. Parfenov VA. [Management of patients with cognitive impairment]. Nevrologiya, neiropsikhiatriya, psikhosomatika – Neurology, Neuropsychiatry, Psychosomatics. 2023. vol. 15(1). pp. 97–102. (In Russ.).
2. Jonell P. et al. Multimodal Capture of Patient Behaviour for Improved Detection of Early Dementia: Clinical Feasibility and Preliminary Results. Frontiers in Computer. Sciences. 2021. vol. 3. DOI: 10.3389/fcomp.2021.642633.
3. Javeed A., Dallora A.L., Berglund J.S., et al. Machine Learning for Dementia Prediction: A Systematic Review and Future Research Directions. Journal of Medical Systems. 2023. vol. 47(17). DOI: 10.1007/s10916-023-01906-7.
4. Lyakso E.E., Frolova O.V., Grechanyj S.V., Matveev Ju.N., Verholjak O.V., Karpov A.A. Golosovoj portret rebenka s tipichnym i atipichnym razvitiem [Vocal profile of a

child with typical and atypical development]. SPb.: Izdatel'sko-poligraficheskaya associaciya vysshih uchebnyh zavedenij. 2020. 204 p. (In Russ.).
5. Velichko A., Karpov A. [Analytical Review of Automatic Systems for Depression Detection by Speech]. Informatics and Automation. 2021. vol. 20(3). pp. 497–529. DOI: 10.15622/ia.2021.3.1. (In Russ.).
6. Bogolepova A.N., et al. [Clinical Guidelines for Cognitive Disorders in Elderly and Older Patients]. Zhurnal nevrologii i psihiatrii im. S.S. Korsakova – S.S. Korsakov Journal of Neurology and Psychiatry. 2022. vol. 121(10-3). pp. 6–137. (In Russ.).
7. Ganguli M., Chang C.C., Snitz B.E., Saxton J.A., Vanderbilt J., Lee C.W. Prevalence of mild cognitive impairment by multiple classifications: The Monongahela-Youghiogheny Healthy Aging Team (MYHAT) project. American Journal of Geriatric Psychiatry. 2010. vol. 18(8). pp. 674–683. DOI: 10.1097/JGP.0b013e3181cdee4f.
8. Larrieu S., Letenneur L, Orgogozo J.M., Fabrigoule C., Amieva H., Le Carret N., Barberger-Gateau P., Dartigues J.F. Incidence and outcome of mild cognitive impairment in a population-based prospective cohort. Neurology. 2002. vol. 59(10). pp. 1594–1599. DOI: 10.1212/01.wnl.0000034176.07159.f8.
9. Roberts R.O, Geda Y.E., Knopman D.S., Cha R.H., Pankratz V.S., Boeve B.F., Ivnik R.J., Tangalos E.G., Petersen R.C., Rocca W.A. The Mayo Clinic Study of Aging: design and sampling, participation, baseline measures and sample characteristics. Neuroepidemiology. 2008. vol. 30(1). pp. 58–69. DOI: 10.1159/000115751.
10. Cherdak M.A. et al. [Prevalence of cognitive impairment in older adults in the Russian Federation]. Zhurnal nevrologii i psihiatrii im. S.S. Korsakova. Specvypuski – S.S. Korsakov Journal of Neurology and Psychiatry. Special issues. 2024. vol. 124(4-2). pp. 5–11. (In Russ.).
11. Wallensten J., Ljunggren G., Nager A., Wachtler C., Bogdanovic N., Petrovic P., Carlsson A.C. Stress, depression, and risk of dementia – a cohort study in the total population between 18 and 65 years old in Region Stockholm. Alzheimer's Research & Therapy. 2023. vol. 15(161). DOI: 10.1186/s13195-023-01308-4.
12. Lokshina A.B., Grishina D.A. [Treatment of noncognitive neuropsychiatric disorders in Alzheimer's disease]. Nevrologiya, neiropsikhiatriya, psikhosomatika – Neurology, Neuropsychiatry, Psychosomatics. 2021. vol. 13(6). pp. 132–138. DOI: 10.14412/2074-2711-2021-6-132-138. (In Russ.).
13. Nestor P.J., Scheltens P., Hodges J.R. Advances in the early detection of Alzheimer's disease. Nature Medicine. 2004. vol. 10. pp. S34–S41. DOI: 10.1038/nrn1433.
14. Botelho C., Abad A., Schultz T., Trancoso I. Speech as a biomarker for disease detection. IEEE Access. 2024. vol. 12. pp. 184487–184508.
15. Yamada Y., Shinkawa K., Ishikawa T., Nishimura M., Nemoto M., Tsukada E., Ota M., Nemoto K., Arai T. Multimodal behavioral analysis for early detection of Alzheimer's disease: A preliminary result: Neuropsychiatry and behavioral neurology/Assessment/Measurement of neuropsychiatric/Behavioral and psychological symptoms. Alzheimer's & Dementia. 2020. vol. 16. DOI: 10.1002/alz.042897.
16. Boschi V., Catricala E., Consonni M., Chesi C., Moro A., Cappa S.F. Connected speech in neurodegenerative language disorders: a review. Frontiers in psychology. 2017. vol. 8. DOI: 10.3389/fpsyg.2017.00269.
17. Hecker P., Steckhan N., Eyben F., Schuller B.W., Arnrich B. Voice analysis for neurological disorder recognition – a systematic review and perspective on emerging trends. Frontiers in Digital Health. 2022. vol. 4. DOI: 10.3389/fdgth.2022.842301.
18. Voleti R., Liss J.M., Berisha V. A review of automated speech and language features for assessment of cognitive and thought disorders. IEEE journal of selected topics in signal processing. 2019. vol. 14(2). pp. 282–298. DOI: 10.1109/jstsp.2019.2952087.

19. Byers A.L., Yaffe K. Depression and risk of developing dementia. Nature Reviews Neurology. 2011. vol. 7(6). pp. 323–331. DOI: 10.1038/nrneurol.2011.60.
20. Braun F., Bayerl S.P., Perez-Toro P.A., Honig F., Lehfeld H., Hillemacher T., Noth E., Bocklet T., Riedhammer K. Classifying Dementia in the Presence of Depression: A Cross-Corpus Study. Proceedings of Interspeech 2023. vol. 2023. pp. 2308–2312. DOI: 10.21437/Interspeech.2023-1997.
21. Jonell P., et al. Multimodal Capture of Patient Behaviour for Improved Detection of Early Dementia: Clinical Feasibility and Preliminary Results. Frontiers in Computer Science. 2021. vol. 3. DOI: 10.3389/fcomp.2021.642633.
22. Lanzi A.M., Saylor A.K., Fromm D., Liu H., MacWhinney B., Cohen M.L. DementiaBank: Theoretical Rationale, Protocol, and Illustrative Analyses. American Journal of Speech-Language Pathology. 2023. vol. 32(2). pp. 426–438. DOI: 10.1044/2022_AJSLP-22-00281.
23. Becker J.T., Boller F., Lopez O.L., Saxton J., McGonigle K.L. The natural history of Alzheimer's disease: description of study cohort and accuracy of diagnosis. Archives of Neurology. 1994. vol. 51(6). pp. 585–594. DOI: 10.1001/archneur.1994.00540180063015.
24. Luz S., Haider F., Fuente S.d.l., Fromm D., MacWhinney B. Alzheimer's Dementia Recognition through Spontaneous Speech: The ADReSS Challenge. Proceedings Interspeech. 2020. pp. 2172–2176. DOI: 10.21437/Interspeech.2020-2571.
25. MacWhinney B. The CHILDES Project: Tools for Analyzing Talk. Child Language Teaching and Therapy. 1992. vol. 8(2). pp. 217–218. DOI: 10.1177/026565909200800211.
26. Luz S., et al. Alzheimer's dementia recognition through spontaneous speech. Frontiers in Computer Science. 2021. vol. 3.
27. Luz S., et al. Connected Speech-Based Cognitive Assessment in Chinese and English. Proceedings Interspeech. 2024. pp. 947–951. DOI: 10.21437/Interspeech.2024-1807.
28. Chen X., Zhang W., Ma Y. Raw Waveform-Based End-to-End Alzheimer's Disease Detection Method. Acta Electronica Sinica. 2023. vol. 51. no. 12. pp. 3582–3590.
29. Псарёва Н.Н. Беглость спонтанной речи как предиктор мягкого когнитивного снижения. ВКР по программе бакалавриата. М.: ВШЭ, 2024.
30. Karakostas A., Briassouli A., Avgerinakis K., Kompatsiaris I., Tsolaki M. The dem@care experiments and datasets: a technical report. arXiv preprint arXiv:1701.01142. 2016.
31. Kishimoto T. et al. PROMPT collaborators. The project for objective measures using computational psychiatry technology (PROMPT): Rationale, design, and methodology. Contemporary Clinical Trials Communications. 2020. vol. 19. DOI: 10.1016/j.conctc.2020.100649.
32. Poor F.F. et al. Prediction of Mild Cognitive Impairment Using a Hybrid Audio-Visual Approach: An I_CONECT Study. Alzheimer's & Dementia Journal. 2023. vol. 19. DOI: 10.1002/alz.074808.
33. Correia J., Teixeira F., Botelho C., Trancoso I., Raj B. The in-the-Wild Speech Medical Corpus. 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2021. pp. 6973–6977. DOI: 10.1109/ICASSP39728.2021.9414230.
34. Orozco J.R. et al. New Spanish speech corpus database for the analysis of people suffering from Parkinson's disease. International Conference on Language Resources and Evaluation (LREC). 2014. pp. 342–347.
35. Gratch J. et al. The Distress Analysis Interview Corpus of human and computer interviews. International Conference on Language Resources and Evaluation (LREC). 2014. pp. 3123–3128.

36. DeVault D. et al. SimSensei kiosk: A virtual human interviewer for healthcare decision support. Proceedings of the 13th International Conference on Autonomous Agents and Multiagent Systems (AAMAS'14). 2014. pp. 1061–1068.
37. Vashkevich M., Rushkevich Yu., Petrovsky A. Bulbar ALS Detection Based on Analysis of Voice Perturbation and Vibrato. Proceedings of inter. conf. Signal Processing: Algorithms, Architectures, Arrangements, and Applications (SPA). 2019. pp. 267–272. DOI: 10.23919/SPA.2019.8936657.
38. Qi X., Zhou Q., Dong J., Bao W. Noninvasive automatic detection of Alzheimer's disease from spontaneous speech: a review. Frontiers in Aging Neuroscience. 2023. vol. 15.
39. Ding K., Chetty M., Noori Hoshyar A., Bhattacharya T., Klein B. Speech based detection of Alzheimer's disease: a survey of AI techniques, datasets and challenges. Artificial Intelligence Review. 2024. vol. 57. no. 12.
40. Babu A. et al. XLS-R: Self-supervised cross-lingual speech representation learning at scale. Proceedings of Interspeech. 2022. pp. 2278–2282. DOI: 10.21437/Interspeech.2022-143.
41. Duan J., Wei F., Li H.D., Liu J. Pre-trained Feature Fusion and Matching for Mild Cognitive Impairment Detection. Proceedings of Interspeech. 2024. pp. 962–966.
42. Favaro A., Cao T., Dehak N., Moro-Velazquez L. Leveraging Universal Speech Representations for Detecting and Assessing the Severity of Mild Cognitive Impairment Across Languages. Proceedings of Interspeech. 2024. pp. 972–976. DOI: 10.21437/Interspeech.2024-2030.
43. Radford A., Kim J.W., Xu T., Brockman G., McLeavey C., Sutskever I. Robust speech recognition via large-scale weak supervision. International conference on machine learning. PMLR 2023. 2023. pp. 28492–28518.
44. Lee B.W., Lee J. LFTK: Handcrafted features in computational linguistics. Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023). 2023. pp. 1–19. DOI: 10.18653/v1/2023.bea-1.1.
45. Hoang B., Pang Y., Dodge H., Zhou J. Translingual Language Markers for Cognitive Assessment from Spontaneous Speech. Proceedings of Interspeech. 2024. 977–981. DOI: 10.21437/Interspeech.2024-1422.
46. Fan A. et al. Beyond English-centric multilingual machine translation. Journal of Machine Learning Research. 2021. vol. 22. no. 107. pp. 1–48.
47. Pérez-Toro P.A., Arias-Vergara T., Klumpp P., Weise T., Schuster M., Noeth E., Orozco-Arroyave J.R., Maier A.. Multilingual Speech and Language Analysis for the Assessment of Mild Cognitive Impairment: Outcomes from the Taukadial Challenge. Proceedings of Interspeech. 2024. pp. 982–986. DOI: 10.21437/Interspeech.2024-2115.
48. Zheng C. et al. Detecting Dementia from Face-Related Features with Automated Computational Methods. Bioengineering (Basel). 2023. vol. 10(7). DOI: 10.3390/bioengineering10070862.
49. Okunishi T., Zheng C., Bouazizi M., Ohtsuki T., Kitazawa M., Horigome T., Kishimoto T. Dementia and MCI Detection Based on Comprehensive Facial Expression Analysis from Videos During Conversation. IEEE Journal of Biomedical and Health Informatics. 2025. vol. 29(5). pp. 3537–3548. DOI: 10.1109/JBHI.2025.3526553.
50. Mu X. et al. Detecting Cognitive Impairment and Psychological Well-being among Older Adults Using Facial, Acoustic, Linguistic, and Cardiovascular Patterns Derived from Remote Conversations. arXiv preprint arXiv:2412.14194. 2024.
51. Jiang Z., Seyedi S., Griner E., Abbasi A., Rad A.B., Kwon H., Cotes R.O., Clifford G.D. Multimodal Mental Health Digital Biomarker Analysis from Remote Interviews

Using Facial, Vocal, Linguistic, and Cardiovascular Patterns. IEEE Journal of Biomedical and Health Informatics. 2024. vol. 28(3). pp. 1680–1691. DOI: 10.1109/JBHI.2024.3352075.
52. Bang J.-U., Han S.-H., Kang B.-O. Alzheimer's Disease recognition from spontaneous speech using large language models. ETRI Journal. 2024. vol. 46. no. 1. pp. 96–105. DOI: 10.4218/etrij.2023-0356.
53. Botelho C. et al. Macro-descriptors for Alzheimer's disease detection using large language models. Proc. Interspeech 2024. 2024. pp. 1975–1979.
54. Haulcy R., Glass J. CLAC: A Speech Corpus of Healthy English Speakers. Proceedings of Interspeech. 2021. pp. 2966–2970. DOI: 10.21437/Interspeech.2021-1810.
55. Dobrovolskii V. Word-Level Coreference Resolution. Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 2021. pp. 7670–7675. DOI: 10.18653/v1/2021.emnlp-main.605.
56. Agarwal R., Melnick L., Frosst N., Zhang X., Lengerich B., Caruana R., Hinton G.E. Neural additive models: Interpretable machine learning with neural nets. Advances in Neural Information Processing Systems. 2021. vol. 34. pp. 4699–4711.
57. Jin D. et al. What Disease Does This Patient Have? A Large-Scale Open Domain Question Answering Dataset from Medical Exams. Applied Sciences. 2021. vol. 11(14). DOI: 10.3390/app11146421.
58. Pal A. et al. MedMCQA: A Large-scale Multi-Subject Multi-Choice Dataset for Medical domain Question Answering. Proceedings of Machine Learning Research. 2022. vol. 174. pp. 248–260.
59. Arora R.K. et al. HealthBench: Evaluating Large Language Models Towards Improved Human Health. arXiv preprint arXiv:2505.08775. 2025.
60. Blinov P., Reshetnikova A., Nesterov A., Zubkova G., Kokh V. RuMedBench: A Russian Medical Language Understanding Benchmark. Conference on Artificial Intelligence in Medicine. 2022. pp. 383–392.
61. Tu T. et al. Towards generalist biomedical AI. NEJM Ai. 2024. vol. 1(3).
62. Driess D. et al. PaLM-E: an embodied multimodal language model. Proceedings of the 40th International Conference on Machine Learning (ICML'23). 2023. vol. 202. pp. 8469–8488.
63. Kim H. et al. Small language models learn enhanced reasoning skills from medical textbooks. NPJ Digital Medicine. 2025. vol. 8(1). DOI: 10.1038/s41746-025-01653-8.
64. Wang Y. et al. Exploiting prompt learning with pre-trained language models for Alzheimer's Disease detection. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). 2023. pp. 1–5. DOI: 10.1109/ICASSP49357.2023.10095993.
65. Balagopalan A., Shkaruta K., Novikova J. Impact of ASR on Alzheimer's Disease Detection: All Errors are Equal, but Deletions are More Equal than Others. Proceedings of the Sixth Workshop on Noisy User-generated Text (W-NUT 2020). 2020. pp. 159–164. DOI: 10.18653/v1/2020.wnut-1.21.
66. Utkin L.V., Konstantinov A.V., Eremenko D.Yu., Zaborovsky V.S., Muliukha V.A. Interpretation methods for machine learning models in the framework of survival analysis with censored data: a brief over view. Computing, Telecommunications and Control. 2024. vol. 17. no. 3. pp. 22–31. DOI: 10.18721/JCSTCS.17302.
67. Alkhatib A. et al. Interpretable Graph Neural Networks for Tabular Data. Proceedings of the 27th European Conference on Artificial Intelligence (ECAI 2024). 2024. vol. 392. pp. 1848–1855.
68. Ren H., Zheng Y., Li C., Jing F., Wang Q., Luo Z., Li D., Liang D., Tang W., Liu L., Cheng W. Using Machine Learning to Predict Cognitive Decline in Older Adults

From the Chinese Longitudinal Healthy Longevity Survey: Model Development and Validation Study. JMIR Aging. 2025. vol. 8. DOI: 10.2196/67437.

69. Karpov A.A., Yusupov R.M. Multimodal Interfaces of Human-Computer Interaction. Herald of the Russian Academy of Sciences. 2018. vol. 88(1). pp. 67–74. DOI: 10.1134/S1019331618010094.

Dolgushin Mikhail — Junior researcher, postgraduate student, Speech and multimodal interfaces laboratory, St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS). Research interests: speech technology, automatic speech recognition, explainable artificial intelligence, computational linguistics. The number of publications — 12. dolgushin.m@iias.spb.su; 39, 14th Line of V.O., 199178, St. Petersburg, Russia; office phone: +7(812)328-0421.

Karpov Alexey — Ph.D., Dr.Sci., Professor, Head of the laboratory, Speech and multimodal interfaces laboratory, St. Petersburg Federal Research Center of the Russian Academy of Sciences (SPC RAS). Research interests: speech technology, automatic speech recognition, audio-visual speech processing, multimodal human-computer interfaces, and computational paralinguistics. The number of publications — 450+. karpov@iias.spb.su; 39, 14th Line of V.O., 199178, St. Petersburg, Russia; office phone: +7(812)328-0421.

Acknowledgements. This research is financially supported by the Russian Science Foundation, project No. 25-11-00319, with the exception of Section 6, which was conducted under the state research No. FFZF-2025-0003.