Выявление признаков депрессии в речи с использованием моделей глубокого обучения

Д. Ш. Дашкин

Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина)

damir.dashkin@gmail.com

Аннотация. **Рост распространённости депрессивных расстройств и ограниченность традиционных методов диагностики, требующих участия специалиста и субъективной оценки, создают потребность в автоматизированных подходах. В работе предлагается метод выявления депрессии по речи с применением глубокого обучения. Для решения задачи разработано комплексное решение, объединяющее сверточную нейронную сеть для извлечения низкоуровневых признаков и предобученную трансформерную модель wav2vec 2.0 для захвата высокоуровневых характеристик речи. Такой подход позволяет преодолеть ограничения существующих методов, основанных на ограниченном наборе признаков, и демонстрирует эффективность автоматизированной диагностики депрессии.**

Ключевые слова: депрессия; обработка речи; глубокое обучение; сверточные нейронные сети; трансформеры; wav2vec 2.0; автоматизированная диагностик

I. ВВЕДЕНИЕ

Диагностика депрессивных состояний традиционно опирается на клинические интервью и самоотчеты пациентов, что вносит элемент субъективности [1]. Речевой сигнал, являясь продуктом сложной нейрофизиологической деятельности, содержит биомаркеры, отражающие психоэмоциональное состояние субъекта [2]. Автоматизирование анализа этих маркеров открывает возможности для объективного скрининга психических расстройств.

Современные исследования в области речевой аналитики разделяются на два направления: использование спектрально-временных представлений (мел-спектрограмм) со сверточными нейронными сетями (сокр. СНС) и применение моделей самообучения, например, таких как Wav2Vec 2.0 [3]. Несмотря на то, что глубокое обучение демонстрирует впечатляющие результаты в детекции депрессивных состояний, большинство существующих решений ограничено использованием признаков лишь одного уровня абстракции [4]. Ситуация осложняется дефицитом качественных клинических данных, что критически затрудняет обучение классических нейронных сетей "с нуля" [5]. В данных условиях максимизация информативности извлекаемых признаков является ключевым фактором повышения точности. В настоящей работе проверяется гипотеза о том, что интеграция разнородных подходов на этапе позднего слияния (англ. late fusion) позволяет модели одновременно верифицировать как локальные акустические аномалии голоса, так и глубокие просодические контексты, обеспечивая синергетический эффект и повышая общую диагностическую значимость системы».

Целью данной работы является разработка комбинированного метода классификации депрессивных состояний, объединяющего спектральные признаки и латентные контекстные репрезентации для повышения точности диагностики на малых выборках.

Для достижения цели были поставлены следующие задачи:

- Выполнить предобработку и нормализацию аудиоданных датасета.
- Реализовать двухпотоковую архитектуру извлечения признаков на основе СНС и модели Wav2Vec 2.0.
- Разработать классификатор на основе метода позднего слияния с регуляризацией.
- Провести оценку эффективности модели на уровне пациента с использованием выбранных метрик.

II. МАТЕРИАЛЫ И МЕТОДЫ

A. *Исходные данные*

Использовались аудиозаписи пациентов датасета MODMA (англ. Multi-modal Open Dataset for Mental-disorder Analysis), собранный в исследовании Ланьчжоуским университетом в 2015 году [6]. Каждому пациенту соответствует метка "MDD" (депрессия) или "Healthy" (здоров). Пример сигнала представлен на рис. 1. Записи предварительно нормализовались и ресемплировались до частоты 16 000 Гц.

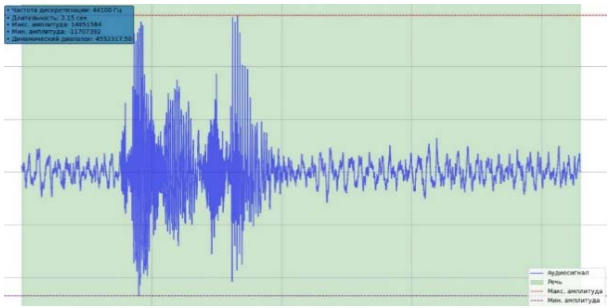

Рис. 1. Визуализация сигнала

B. *Инструменты*

Разработка велась на языке Python 3.10 [7] с использованием фреймворка глубокого обучения PyTorch [8]. Для обработки аудиосигналов применялись библиотеки Librosa [9] и Torchaudio [10]. Высокоуровневые признаки извлекались с помощью предобученной модели «facebook/wav2vec2-base-960h» из репозитория Hugging Face [11].

C. *Метрики качества*

Для оценки были выбраны метрики доли правильных ответов (англ. Accuracy), точности классификации положительного класса (англ. Precision), чувствительности (англ. Recall) и значение F1-меры (англ. F1-score). Выбор обусловлен необходимостью минимизации пропуска больных пациентов (высокое значение метрики полноты) и исключения ложноположительных диагнозов (высокое значение метрики положительных ответов). Согласно [4, 5], F1-score является наиболее репрезентативной метрикой при работе с медицинскими данными малого объема, так как представляет собой гармоническое среднее между точностью и полнотой.

D. *Методы предобработки данных*

Процесс подготовки первичных аудиосигналов к анализу представлял собой многоступенчатую процедуру трансформации, направленную на стандартизацию входных параметров и искусственное расширение обучающей выборки. На начальном этапе каждый аудиофайл подвергался передискретизации до частоты 16 кГц, что является стандартом для работы современных моделей речевого анализа, и преобразовывался в монофонический формат. Для компенсации различий в условиях записи и аппаратном обеспечении применялась амплитудная нормализация (по пиковому значению, приводящая динамический диапазон всех сигналов к единому масштабу. Поскольку исходные записи имели значительную длительность, была реализована процедура сегментации, разделяющая сигнал на фрагменты фиксированной длины. Это позволило кратно увеличить количество обучающих примеров и обеспечить стабильную работу механизма мажоритарного голосования на этапе тестирования.

Для извлечения частотно-временных характеристик сигнала выполнялось спектральное преобразование, включающее вычисление 80-канальных мел-спектрограмм (англ. Mel-Spectrograms) с последующим логарифмированием. Полученные признаки проходили процедуру стандартизации, что позволило модели сверточной нейронной сети эффективно распознавать паттерны в частотной области вне зависимости от индивидуальной громкости диктора. Дополнительно в последовательность обработки были интегрированы методы аугментации данных, применявшиеся только к тренировочному набору. Они включали наложение случайного шума, временную деформацию и изменение высоты тона. Совокупное использование данных методов позволило предотвратить переобучение нейронной сети на специфических характеристиках дикторов и сфокусировать внимание классификатора на интонационных биомаркерах, характерных для аффективных состояний.

E. *Процедура обучения*

Процесс экспериментального исследования был организован на основе системы инкрементального кэширования признаков, что позволило оптимизировать вычислительные затраты. На начальном этапе производилась проверка существующего локального хранилища на предмет наличия данных по списку идентификаторов из таблицы информации о пациентах. При обнаружении отсутствующих записей система автоматически инициировала процесс извлечения признаков аудиофайлы загружались с частотой 16 кГц, после чего для каждого фрагмента вычислялись спектральные и контекстные характеристики. Обновленный массив данных, включающий 80-канальные мел-спектрограммы, векторы модели Wav2Vec 2.0 и соответствующие им клинические метки, сохранялся в файл формата Pickle.

Для проведения эксперимента сформированная выборка разделялась на обучающий и тестовый наборы в пропорции 75% к 25% с применением стратификации (англ. Stratification), что исключало попадание данных одного пациента в разные группы. Обучение нейронной сети проводилось в течение 40 эпох с использованием алгоритма адаптивной оценки момента (англ. Adam) с начальной скоростью обучения 2e-4 и коэффициентом регуляризации весов (англ. Weight Decay) 1e-3. Для компенсации дисбаланса классов в функцию потерь в виде бинарной перекрестной энтропии с логитами (англ. Binary Cross Entropy with Logits) был введен весовой коэффициент 1.2. Итоговое диагностическое решение принималось на уровне пациента методом мажоритарного голосования: диагноз выставлялся, если более 50% сегментов записи классифицировались как положительные при пороге вероятности 0.45. Параметры лучшей модели сохранялись при достижении максимального показателя доли правильных ответов на тестовом наборе.

F. *Методы анализа и архитектура модели*

Для решения задачи классификации была разработана мультимодальная нейронная сеть, реализующая стратегию позднего слияния признаков, извлеченных из различных представлений аудиосигнала. Архитектура модели спроектирована таким образом, чтобы одновременно анализировать как низкоуровневые частотно-временные характеристики, так и высокоуровневые семантические признаки речи. Модель состоит из двух независимых ветвей обработки, выходы которых объединяются для принятия итогового решения.

Первая ветвь (спектральная) предназначена для обработки 80-канальных лог-мел спектрограмм. Входные данные подаются в виде тензора с одним каналом, который проходит через слой двумерной свертки (англ. 2D Convolution) с 16 фильтрами и ядром размером 3x3 при наличии нулевого дополнения (англ. Padding). Сверточный слой позволяет извлекать локальные паттерны интенсивности звука на разных частотах. Сразу после свертки применяется слой пакетной нормализации (англ. Batch Normalization) для стабилизации распределения активаций и ускорения сходимости, а также нелинейная функция активации линейного выпрямителя (англ. Rectified Linear Unit, ReLU). Для приведения признаковых карт произвольной длительности к единому размеру используется слой глобального адаптивного пулинга (англ. Adaptive Average Pooling), который вычисляет среднее значение по всей временной и частотной осям. Результирующий вектор проходит через полносвязный слой (англ. Linear Layer), проецирующий данные в 64-мерное пространство с последующей активацией ReLU.

Вторая ветвь (контекстная) отвечает за обработку 768-мерных векторных представлений (англ. Embeddings), полученных из предобученной модели

Wav2Vec 2.0 (архитектура base-960h). Поскольку данные векторы уже содержат глубокую информацию о структуре речи, они обрабатываются специализированным проекционным блоком. Этот блок состоит из полносвязного слоя, преобразующего 768 признаков в 64-мерный вектор, функции активации ReLU и слоя регуляризации путем исключения нейронов (англ. Dropout) с коэффициентом 0.5. Использование Dropout на данном этапе критически важно для предотвращения переобучения модели на специфических особенностях дикторов из обучающей выборки.

Интегральный классификатор выполняет слияние данных путем сцепления двух 64-мерных векторов от обеих ветвей, в результате чего формируется единый вектор признаков размерностью 128. Этот вектор подается на вход финальной решающей сети, которая имеет двухслойную структуру. Скрытый слой классификатора содержит 32 нейрона с активацией ReLU, за которым следует еще один слой Dropout (0.5) для дополнительной устойчивости. Выходной слой состоит из одного нейрона, формирующего итоговое логит-значение. Данная многоуровневая структура позволяет модели выявлять сложные нелинейные зависимости между спектральными аномалиями голоса и контекстными характеристиками речи, обеспечивая высокую точность идентификации целевых состояний.

III. Результаты

Апробация разработанного мультимодального метода на независимой тестовой выборке, включающей 13 пациентов, позволила получить количественные оценки эффективности предложенной архитектуры. Итоговые показатели рассчитывались на уровне субъектов исследования после процедуры мажоритарного голосования по всем аудиосегментам каждого участника. Результаты представлены в таблице 1.

Согласно полученным данным, доля правильных ответов составила 0.77 (77%). Особое внимание заслуживает показатель чувствительности для целевой группы пациентов с депрессивным расстройством (метка MDD), который достиг максимального значения 1.00. Это свидетельствует о том, что система успешно идентифицировала всех пациентов с подтвержденным диагнозом, не допустив ни одного случая пропуска заболевания. Одновременно с этим, показатель прогностической ценности или по-другому точности классификации положительного класса для контрольной группы здоровых участников также составил 1.00, что подтверждает отсутствие ложноположительных срабатываний среди лиц без патологии. Значение F1-меры в макро-усреднении составило 0.76, что подтверждает сбалансированность модели. Таким образом, система продемонстрировала исключительную чувствительность к биомаркерам депрессии в голосе, обеспечив полное покрытие больных пациентов при сохранении высокого общего уровня точности.

ТАБЛИЦА I.

Метрика	Значение
Доля правильных ответов	0.77
Точность положительного класса	0.67
Чувствительность	1.00
Оценка F1-меры	0.76

IV. Обсуждение

Результаты экспериментального исследования подтверждают гипотезу о том, что интеграция разнородных признаков в рамках единой нейросетевой архитектуры позволяет эффективно решать задачу автоматизированной диагностики даже в условиях ограниченной выборки (51 человек). Достижение предельного значения полноты (1.00) указывает на то, что сцепление признаков, извлеченных сверточной сетью, и глубоких векторных представлений позволяет модели фиксировать специфические акустические и лингвистические паттерны, которые зачастую недоступны при использовании изолированных спектральных методов.

Некоторый разрыв между чувствительностью и точностью для положительного класса (Precision MDD составил 0.67) объясняется склонностью модели к гипердиагностике. В контексте медицинской информатики такая особенность является допустимой и даже предпочтительной для систем первичного скрининга, где критически важно не пропустить потенциального пациента. Важнейшим фактором стабилизации работы системы стало внедрение механизма агрегации решений на уровне пациента. Переход от анализа разрозненных фрагментов к итоговому голосованию позволил нивелировать влияние случайных шумов и артефактов в отдельных записях, что существенно повысило надежность диагностики по сравнению с традиционными методами пофрагментной классификации.

Несмотря на достигнутые высокие показатели чувствительности, данное исследование имеет ряд ограничений, обусловленных спецификой исходных данных и вычислительной архитектуры. Относительно малый объем выборки (51 участник) накладывает определенные рамки на обобщающую способность модели, что в будущем потребует проверки алгоритмов на более масштабных и разнообразных наборах данных, включающих различные возрастные и социальные группы. Кроме того, текущая реализация мажоритарного голосования присваивает всем аудиофрагментам пациента равные веса, однако в реальной речи информативность различных сегментов может варьироваться. В связи с этим, перспективным направлением развития системы является внедрение механизмов внимания, которые позволят нейронной сети автоматически выделять наиболее диагностически значимые участки аудиозаписи, игнорируя паузы или неинформативные шумы.

Дальнейшее совершенствование метода также может быть связано с переходом от использования фиксированных векторных представлений к процедуре тонкой настройки всей архитектуры Wav2Vec 2.0 под специфические задачи эмоционально-речевого анализа. Это позволит адаптировать предобученные веса трансформерной модели непосредственно под акустические маркеры депрессивных состояний. Также представляется целесообразным расширение мультимодального подхода путем интеграции текстовых транскрипций речи, полученных с помощью систем автоматического распознавания речи, что даст возможность анализировать не только акустику, но и семантическое содержание высказываний пациента. Подобная синергия методов обработки сигналов и

естественного языка способна значительно повысить точность классификации и снизить уровень ложноположительных срабатываний.

V. Заключение

В ходе выполнения данной работы был реализован полный цикл разработки системы поддержки принятия врачебных решений в области цифровой психиатрии. В рамках исследования были успешно решены ключевые задачи: разработана методика многоэтапной предобработки аудиосигналов с применением инкрементального кэширования, спроектирована гибридная архитектура на базе сверточных слоев и трансформерных моделей Wav2Vec 2.0, а также внедрен алгоритм мажоритарного голосования для оценки состояния на уровне субъекта.

Практическая значимость работы подтверждается достигнутыми метриками качества: точностью 0.77 и F-мерой 0.76. Полученные результаты доказывают перспективность использования современных архитектур глубокого обучения для объективизации процесса диагностики аффективных расстройств. Дальнейшее развитие проекта может быть направлено на расширение обучающей выборки и интеграцию дополнительных модальностей, что позволит еще больше снизить вероятность ложноположительных заключений при сохранении высокой чувствительности системы.

Список литературы

[1] Депрессивное расстройство (депрессия) [Электронный ресурс] // Всемирная организация здравоохранения. – URL: https://www.who.int/ru/news-room/fact-sheets/detail/depression

[2] Major Depressive Disorder [Электронный ресурс] // National Library of Medicine. – URL: https://www.ncbi.nlm.nih.gov/books/NBK559078

[3] Aleagha D.M., Zohari P., Chehreghani M.H. AI Models for Depressive Disorder Detection and Diagnosis: A Review [Электронный ресурс]. 2025. – URL: https://arxiv.org/abs/2508.12022

[4] Mao K., Wu Y., Chen J. A systematic review on automated clinical depression diagnosis [Электронный ресурс] // npj Mental Health Research. 2023. – URL: https://www.nature.com/articles/s44184-023-00040-z

[5] Zhao Z., Li J., Wang Z. [и др.]. Automatic depression recognition by intelligent speech signal processing: A systematic survey [Электронный ресурс] // Cognitive Computation and Systems. 2023. – URL: https://ietresearch.onlinelibrary.wiley.com/doi/epdf/10.1049/cit2.12113

[6] Multi-modal Open Dataset for Mental-disorder Analysis (MODMA) [Электронный ресурс] // UK Data Service ReShare. – URL: https://reshare.ukdataservice.ac.uk/854301

[7] Python [Электронный ресурс]. – URL: https://docs.python.org/3.10/

[8] PyTorch [Электронный ресурс]. – URL: https://pytorch.org/

[9] Librosa [Электронный ресурс]. – URL: https://librosa.org/doc/latest/index.html

[10] Torchaudio [Электронный ресурс]. – URL: https://docs.pytorch.org/audio/stable/index.html

[11] Wav2Vec2-Base-960h [Электронный ресурс]. – URL: https://huggingface.co/facebook/wav2vec2-base-960h