Сравнительный анализ и интеграция речевых представлений для выявления депрессии

Д. Ш. Дашкин

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

damir.dashkin@gmail.com

Аннотация. **В работе рассматривается задача автоматического выявления депрессивных состояний по речевым данным. Исследуются три подхода к представлению сигнала: спектральные признаки, обрабатываемые сверточной нейронной сетью, высокоуровневые представления, извлекаемые предобученной моделью wav2vec 2.0, а также их объединение. Проведен сравнительный анализ указанных подходов и оценено влияние объединения признаков на качество классификации.**

Ключевые слова: депрессия; анализ речи; глубокое обучение; сверточные нейронные сети; wav2vec 2.0; объединение признаков

I. Введение

Депрессивные расстройства являются одной из наиболее распространенных форм психических заболеваний и оказывают значительное влияние на качество жизни пациентов [1]. Традиционные методы диагностики, основанные на клинических интервью и самооценке, характеризуются субъективностью и требуют участия квалифицированных специалистов. В связи с этим возрастает интерес к разработке автоматизированных методов анализа, способных выявлять признаки депрессии на основе объективных данных.

Речевой сигнал представляет собой информативный источник, отражающий когнитивные и эмоциональные процессы [2]. В последние годы сформировались два основных подхода к анализу речи в задачах диагностики: использование спектральных признаков (например, мел-спектрограмм) и применение самообучающихся моделей, таких как wav2vec 2.0 [3]. Первый подход позволяет выявлять локальные акустические особенности сигнала, тогда как второй — извлекать высокоуровневые контекстные представления.

Несмотря на успешность каждого из подходов, их сравнительный анализ и исследование совместного использования остаются актуальными задачами [4]. В данной работе проверяется гипотеза о том, что объединение спектральных и контекстных признаков позволяет повысить качество классификации депрессивных состояний.

Целью работы является проведение сравнительного анализа различных представлений речевого сигнала и оценка эффективности их интеграции.

Для достижения цели были поставлены следующие задачи:

- реализовать извлечение низкоуровневых признаков с помощью сверточной нейронной сети;
- реализовать извлечение представлений wav2vec 2.0;
- разработать модели классификации для каждого типа признаков;
- реализовать комбинированную модель с объединением признаков;
- провести экспериментальное сравнение подходов.

II. Материалы и методы

A. *Исходные данные*

В работе использовался датасет MODMA (Multi-modal Open Dataset for Mental-disorder Analysis), содержащий аудиозаписи пациентов с диагнозом депрессии (метка MDD) и здоровых участников (метка Healthy) [5]. Все записи были приведены к частоте дискретизации 16 кГц. Пример сигнала до ресемплирования представлен на рис. 1.

Разделение выборки осуществлялось на уровне пациентов с использованием стратифицированного разбиения в пропорции 75% (обучение) и 25% (тестирование), что исключало попадание данных одного пациента в обе выборки.

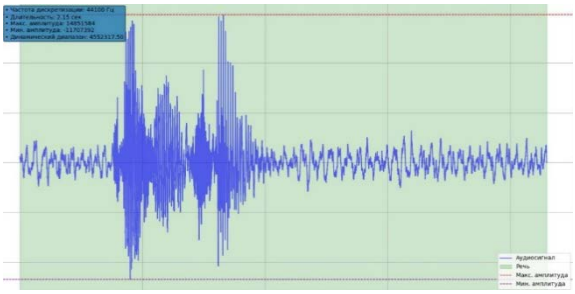

Рис. 1. Визуализация сигнала

B. *Инструменты*

Разработка программной реализации проводилась на языке Python 3.10 [6] с использованием фреймворка глубокого обучения PyTorch [7]. Для обработки аудиосигналов применялась библиотека Librosa [8], обеспечивающая загрузку и преобразование звуковых данных, а также вычисление спектральных признаков.

Для формирования обучающих и тестовых выборок, а также оценки качества моделей использовались инструменты библиотеки Scikit-learn [9], включая

функции стратифицированного разбиения данных и расчета метрик классификации.

Высокоуровневые признаки извлекались с использованием предобученной модели «facebook/wav2vec2-base-960h» из библиотеки Hugging Face Transformers [10]. Модель применялась в режиме извлечения признаков без дополнительной дообучения, а итоговое представление формировалось путем усреднения скрытых состояний.

C. Метрики качества

Для оценки качества классификации использовались следующие метрики: доля правильных ответов, точность (англ. Accuracy), точность положительного класса (англ. Precision), чувствительность (англ. Recall) и F1-мера (англ. F1-score).

Выбор данных метрик обусловлен спецификой задачи диагностики депрессивных состояний, где критически важно обеспечить баланс между выявлением всех пациентов с заболеванием (высокая полнота) и минимизацией ложноположительных решений (высокая точность).

Дополнительно использовалась метрика площади под ROC-кривой (англ. ROC-AUC), позволяющая оценить качество ранжирования объектов по вероятности принадлежности к классу депрессии.

Согласно [4, 5], F1-мера является наиболее репрезентативной метрикой при работе с медицинскими данными малого объема, так как представляет собой гармоническое среднее между точностью и полнотой и позволяет учитывать, как ошибки первого, так и второго рода.

D. Методы предобработки данных

Подготовка аудиоданных к анализу включала последовательность преобразований, направленных на приведение сигналов к единому формату и извлечение информативных признаков.

На первом этапе аудиофайлы загружались и автоматически приводились к частоте дискретизации 16 кГц с использованием библиотеки Librosa. Дополнительная нормализация амплитуды или сегментация сигналов не применялись.

Для извлечения спектральных признаков вычислялись мел-спектрограммы (англ. Mel-Spectrograms) с 80 фильтрами. Полученные спектрограммы преобразовывались в логарифмическую шкалу и далее нормализовывались путем вычитания среднего значения и деления на стандартное отклонение. Итоговое представление формировалось в виде тензора с одним каналом.

Высокоуровневые признаки извлекались с использованием предобученной модели wav2vec 2.0. Аудиосигнал подавался на вход модели, после чего из последнего скрытого слоя извлекались признаки, усредненные по временной оси, что позволяло получить фиксированное векторное представление размерности 768.

Для повышения эффективности вычислений реализован механизм кэширования признаков. При первом запуске выполнялось извлечение спектральных и wav2vec-признаков для всех аудиофайлов, после чего они сохранялись в файл формата Pickle. При последующих запусках данные загружались из кэша без повторной обработки.

E. Процедура обучения

Экспериментальное исследование проводилось с использованием предварительно сформированного набора признаков, включающего мел-спектрограммы, wav2vec-представления и метки классов.

Разделение данных на обучающую и тестовую выборки выполнялось на уровне пациентов в пропорции 75% к 25% с использованием стратификации (англ. Stratification), что исключало попадание данных одного пациента в обе выборки.

Обучение моделей осуществлялось на уровне отдельных аудиофайлов. Для формирования батчей использовался DataLoader, при этом спектрограммы внутри батча приводились к одинаковой длине с помощью дополнения нулями.

Обучение выполнялось в течение 15 эпох с использованием оптимизатора адаптивной оценки момента (англ. Adaptive Moment Estimation, сокр. Adam) с начальной скоростью обучения 2e-4. В качестве функции потерь применялась бинарная кросс-энтропия с логитами (англ. Binary Cross Entropy with Logits).

Оценка качества модели проводилась на уровне пациента. Для каждого пациента вычислялись вероятности принадлежности к классу депрессии по всем его аудиофайлам, после чего рассчитывалось их среднее значение. Итоговое решение принималось путем пороговой классификации со значением порога 0.5.

F. Методы анализа и архитектура модели

Для решения задачи классификации была реализована нейронная сеть с возможностью работы в трех режимах: спектральном (англ. CNN), контекстном (англ. W2V) и комбинированном (англ. Fusion). Такой подход позволяет в рамках единой архитектуры провести корректный сравнительный анализ различных типов признаков, а также оценить эффект их интеграции.

Спектральная ветвь модели представляет собой сверточную нейронную сеть, принимающую на вход мел-спектрограммы. Архитектура включает несколько последовательных слоев двумерной свертки с нелинейной функцией активации (англ. Rectified Linear Unit, ReLU), сгруппированных в блоки. После первых двух сверток применяется операция подвыборки (англ. MaxPooling), позволяющая уменьшить размерность признаковых карт и выделить более устойчивые паттерны. Далее аналогичная структура повторяется с увеличением числа каналов (16, 32, 64), что позволяет постепенно переходить от локальных акустических особенностей к более абстрактным представлениям сигнала. На завершающем этапе применяется адаптивный глобальный пулинг, приводящий данные к фиксированной размерности независимо от длины входного сигнала. Полученный вектор преобразуется полносвязным слоем в 128-мерное пространство признаков.

Контекстная ветвь обрабатывает 768-мерные векторные представления, извлекаемые из

предобученной модели wav2vec 2.0. Поскольку данные признаки уже содержат высокоуровневую информацию о структуре речи, их обработка сводится к линейному преобразованию в пространство меньшей размерности. Для этого используется полносвязный слой с нелинейной функцией активации ReLU, формирующий 128-мерный вектор. Такое преобразование обеспечивает согласование размерностей с выходом спектральной ветви и позволяет корректно объединять признаки.

В зависимости от выбранного режима работы используются разные варианты классификации:

- в режиме CNN классификация выполняется только на основе спектральных признаков.
- в режиме W2V используется только wav2vec-представление.
- в режиме Fusion выполняется объединение признаков путем конкатенации двух 128-мерных векторов, формируя итоговый вектор размерности 256.

Полученный вектор подается на линейный слой, формирующий итоговое логит-значение.

Использование единой архитектуры с переключаемыми режимами позволяет проводить корректное сравнение подходов при одинаковых условиях обучения. Это исключает влияние различий в классификаторе и делает возможной интерпретацию результатов исключительно с точки зрения используемых признаков.

Дополнительно, выбранная стратегия объединения признаков представляет собой прямую конкатенацию без применения сложных механизмов взаимодействия. Это позволяет изолированно оценить вклад каждого типа представлений и выявить наличие синергетического эффекта при их совместном использовании.

III. Результаты

Апробация разработанного метода, ориентированного на сравнительный анализ и интеграцию речевых представлений, была проведена на независимой тестовой выборке, включающей 13 пациентов. В рамках эксперимента оценивались три режима работы модели: спектральный (CNN), контекстный (W2V) и комбинированный (Fusion). Итоговые показатели рассчитывались на уровне субъектов исследования путем усреднения вероятностей по всем аудиозаписям каждого пациента с последующей пороговой классификацией. Результаты представлены в табл. 1.

Согласно полученным данным, наилучшие результаты были достигнуты в режиме Fusion, где точность положительного класса составила 0.77 (77%). Для данного режима точность классификации для класса Healthy составила 0.75, а для класса MDD — 0.80. Чувствительность для здоровых участников достигла 0.86, тогда как для пациентов с депрессивным расстройством — 0.67. Значение F1-меры в макро-усреднении составило 0.76, что свидетельствует о сбалансированности модели.

При этом отдельный анализ показал, что использование только спектральных или только контекстных признаков приводит к одинаковому уровню точности (0.62), однако различается по показателю ROC-AUC: 0.50 для CNN и 0.71 для W2V. Это указывает на то, что высокоуровневые представления wav2vec 2.0 лучше отражают разделимость классов, тогда как их изолированное использование недостаточно для повышения итоговой точности классификации. Таким образом, объединение признаков позволяет повысить качество модели по сравнению с каждым из подходов в отдельности. С более подробными результатами метрик можно ознакомиться в табл. 1.

ТАБЛИЦА I.

Показатель\Модель	CNN	W2V	Fusion
Доля правильных ответов	0.62	0.62	0.77
Точность положительного класса	0.57	0.57	0.8
Чувствительность	0.67	0.67	0.67
Оценка F1-меры (макро)	0.62	0.62	0.76
Значение ROC-AUC	0.5	0.71	0.6

IV. Обсуждение

Результаты экспериментального исследования подтверждают гипотезу о том, что интеграция разнородных речевых представлений в рамках единой архитектуры позволяет повысить качество автоматической диагностики по сравнению с использованием отдельных типов признаков. Существенный прирост точности при переходе к режиму Fusion демонстрирует наличие дополнительной информации, извлекаемой при совместном анализе спектральных и контекстных характеристик речи.

Сравнительный анализ показал, что изолированное использование CNN и W2V приводит к сопоставимым значениям точности (0.62), однако различия в ROC-AUC указывают на неодинаковую природу извлекаемых признаков. Контекстные представления wav2vec 2.0 обеспечивают лучшую разделимость классов, в то время как спектральные признаки отражают локальные акустические особенности сигнала. Их объединение позволяет компенсировать ограничения каждого подхода и формировать более устойчивое представление.

Отдельного внимания заслуживает тот факт, что итоговое решение принималось на уровне пациента путем усреднения вероятностей по всем его записям. Такой подход позволил снизить влияние вариативности отдельных аудиофрагментов и повысить устойчивость модели к шумам и нестабильности речи. В отличие от пофрагментной классификации, данный метод обеспечивает более надежную оценку состояния пациента в целом.

Несмотря на улучшение качества в режиме Fusion, показатели чувствительности и точности остаются несбалансированными для отдельных классов, что может свидетельствовать о чувствительности модели к особенностям выборки. Ограниченный объем данных (51 участник) накладывает ограничения на обобщающую способность модели и требует дальнейшей валидации на более масштабных наборах данных.

Перспективным направлением развития является расширение подхода за счет более сложных методов интеграции признаков. В текущей работе используется

простая конкатенация, однако в дальнейшем возможно применение механизмов, позволяющих учитывать взаимосвязи между различными типами признаков. Кроме того, потенциальное улучшение может быть достигнуто за счет тонкой настройки модели wav2vec 2.0 под задачу выявления депрессивных состояний, а также за счет использования дополнительной информации, например, текстовых транскрипций речи.

V. Заключение

В ходе выполнения данной работы был реализован метод анализа речевых сигналов для выявления депрессивных состояний, основанный на сравнении и интеграции различных типов признаков. Были рассмотрены три режима работы модели: спектральный, контекстный и комбинированный, что позволило провести всесторонний анализ их эффективности в рамках единой архитектуры.

Экспериментальные результаты показали, что наилучшее качество достигается при объединении признаков: точность классификации составила 0.77, а значение F1-меры — 0.76. Это подтверждает целесообразность использования мультимодального подхода при анализе речевых данных. Таким образом, были выполнены все поставленные задачи.

Полученные результаты демонстрируют потенциал применения методов глубокого обучения для задач цифровой психиатрии и автоматизированного скрининга депрессивных расстройств. Дальнейшее развитие работы может быть связано с увеличением объема данных, улучшением методов интеграции признаков и адаптацией предобученных моделей к специфике рассматриваемой задачи.

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

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

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

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

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

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