УДК 004.81
DOI:10.25729/ESI.2025.40.4.005

Обнаружение мультимодальной депрессии с использованием многопотоковой модели Mood Insight Encoder (MMIE)

Фироз Неда[1], Берестнева Ольга Григорьевна[2], Аксенов Сергей Владимирович[2]

[1]Томский государственный университет,
Россия, Томск, *nedafiroz1910@gmail.com*
[2]Томский политехнический университет,
Россия, Томск

Аннотация. Глобальный всплеск распространенности депрессии, которая характеризуется стойким чувством печали, незаинтересованности и снижением функциональности, подчеркивает недостатки преобладающих парадигм диагностики и лечения. Это подчеркивает настоятельную потребность в усиленных вмешательствах, учитывая недостатки, присущие традиционным подходам к диагностике депрессии. Недавние достижения в области применения искусственного интеллекта вызвали растущий интерес к разработке автоматизированных систем диагностики депрессии среди специалистов по эмоциональным вычислениям. Появление больших языковых моделей, таких, как BERT и его производные, для выявления депрессии на основе текста демонстрирует необходимость мультимодальных подходов, которые объединяют текстовые и аудиоданные модальности для достижения более точной диагностики. В этой работе авторы исследовали возможности существующих больших языковых моделей и предложили многопоточную модель Multi-Stream Mood Insight Encoder (MMIE). MMIE разработана для беспрепятственного использования интегрированных потоков текстовых и аудиоданных с функциями обработки с помощью кодировщика Reformer. В рамках этой концепции в кодировщик Reformer были включены такие лингвистические особенности, как абсолютистские слова и местоимения первого лица. Такой целостный подход способствовал всестороннему анализу настроения и эмоционального состояния человека. Эксперименты показали, что языковая модель ClinicalBERT превосходит по производительности предложенную модель бинарной классификации депрессии. Впоследствии для диагностики депрессии были использованы значения сигмовидной мышцы модели Реформер. Используя предложенную модель, были проведены эксперименты с набором данных DAIC-WOZ. Результаты показали значительные улучшения, продемонстрировав F1 0,9538 при классификации, в то время как MAE 3,42 и RMSE 4,64 при регрессии по сравнению с самыми современными методами. Эти результаты демонстрируют эффективность предложенной модели в облегчении диагностики депрессии.

Ключевые слова: аудио, клинический анализ, выявление депрессии, LLMs, Реформер, MMIE

Цитирование: Фироз Н. Обнаружение мультимодальной депрессии с использованием многопотоковой модели Mood Insight Encoder (MMIE) / Н. Фироз, О.Г. Берестнева, С.В. Аксенов // Информационные и математические технологии в науке и управлении, 2025. – № 4(40). – С. 52-77. – DOI:10.25729/ESI.2025.40.4.005.

Введение. В целом, каждый пятый взрослый страдает тяжелым депрессивным расстройством (МДД), причем его распространенность растет в последнее десятилетие из-за проблем общественного здравоохранения и снижения производительности [1-2]. Серьезное депрессивное расстройство (МДД), или депрессия, определяется грустным, опустошенным или раздражительным настроением, наряду с соматическими и когнитивными изменениями, влияющими на функционирование [3]. Многочисленные демографические факторы влияют на это сложное заболевание, причины которого лучше распознаны, но существующие методы лечения недостаточны [4-5]. Автоматизация в общественном здравоохранении необходима для постановки сложных диагнозов, а диагностика на базе искусственного интеллекта обещает раннее выявление симптомов и предотвращение самоубийств [6-7]. В то время как глубокое обучение превосходно работает с большими наборами данных, традиционные модели остаются эффективными для наборов данных меньшего размера [8-10].

В ответ на растущий интерес к изучению анализа эмоций и депрессии появились базы данных, такие, как AVEC (2016, 2017, 2019). В соответствии с авторским исследованием взаимосвязи языка и психического здоровья, в различных исследованиях [11-15]

использовалась обработка естественного языка (НЛП) для прогнозирования психических расстройств. Используемый набор функций оказывает существенное влияние на эффективность моделей глубокого обучения [16].

Представление двунаправленного кодера с помощью transformers (BERT) обладает мощными возможностями извлечения признаков [17], обеспечивая самые современные результаты для задач обработки естественного языка (NLP). Предварительно обученные модели, такие, как MentalBERT, MentalRoBERTa [18], PsychBERT [19] и ClinicalBERT [20], используют общедоступные данные в соответствующих областях для извлечения текстовых признаков, широко признанных, наряду с BERT, в мультимодальной классификации эмоций [21-23]. Однако, за исключением нескольких исследований [11, 24], они не были широко использованы для этой цели. Используя наборы данных, относящиеся к психическому здоровью, эти модели были адаптированы к предметной области или доработаны специально для приложений здравоохранения.

Исследования физиологических аспектов депрессии выявили заметные различия в голосовых движениях депрессивных и здоровых людей [25-26]. Как правило, необработанные аудиоданные используются для создания изображений спектрограмм и низкоуровневых дескрипторов (LLD) для прогнозирования уровня снижения, таких показателей, как громкость, энтропия, дисперсия, асимметрия и кепстральные коэффициенты малой частоты (MFCC). [27-28].

Хотя текстовые данные дают представление о депрессивных состояниях [11], полагаться исключительно на них недостаточно, поскольку зашумленные или неполные данные могут повлиять на производительность модели. Чтобы устранить эту проблему, мультимодальная идентификация депрессии использует текстовые и слуховые данные для идентификации аффективных состояний [29]. Таким образом, в этом исследовании используются функции COVAREP [30] и FORMANT [31] для аудиоданных из набора данных DAIC-WOZ [32] и NLP для текстовых данных для диагностики депрессии, повышая точность благодаря анализу эталонного корпуса.

Толбелл предложил тщательное изучение языковых черт, характерных для депрессии, которые были представлены в 26 публикациях [33]. Меньшее внимание уделялось лингвистическим особенностям, таким, как абсолютистские слова [34] и местоимения от первого лица [33], которые указывают на их использование участниками с депрессией.

Большинство методов когнитивной терапии тревоги и депрессии рассматривают абсолютистское мышление как когнитивную ошибку. Слова или фразы, которые подразумевают абсолютную уверенность или определенность, известны, как абсолютистские слова. Они не дают большой свободы для интерпретации или различных толкований. Эти фразы часто передают идею о том, что существует только один правильный ответ или точка зрения, игнорируя сложность или противоположные идеи. Среди терминов, подпадающих под эту категорию, – "всегда", "никогда", "полностью", "тотально", "навсегда", "невозможно" и "неизбежно". Эти фразы иногда могут приводить к догматическим рассуждениям или резким точкам зрения, игнорирующим богатство и разнообразие реальных обстоятельств [34].

Предыдущие исследования в основном были сосредоточены на методах обработки естественного языка (NLP) для извлечения признаков при обнаружении депрессии на основе текста [35] и разработке моделей глубокого обучения. Однако в этих исследовательских работах не был полностью использован потенциал аудиомодальности в сочетании с языковыми моделями на основе трансформеров (LLM), а именно MentalBERT, MentalRoBERT, PsyhBERT and ClinicalBERT [36], для тщательного выделения признаков и прогнозирования мультимодальной депрессии.

Разработка универсально оптимальной модели для автоматизированной диагностики депрессии сопряжена с трудностями [37]. Основанные на тексте алгоритмы обнаружения депрессий обычно отдают предпочтение надежной модели глубокого обучения LSTM, которая устраняет проблемы исчезновения градиента, связанные с RNN, благодаря своей простой структуре и методам стробирования [38-40]. Однако последовательная сетевая структура LSTM, где каждый прямой проход зависит от результата предыдущего временного шага, создает проблемы с эффективностью для параллельной обработки [41]. Преобразователи, с другой стороны, обеспечивают параллельную обработку с использованием позиционного кодирования, устраняя недостатки LSTM в обработке последовательного ввода [42]. Техника многоголового самонаблюдения помогает распознавать важные элементы в данных [41]. Преобразователь, существенно более быстрый и экономичный по сравнению с моделями трансформеров, работает сравнительно хорошо с ними при обработке длинных последовательностей [43].

Более того, в недавних исследованиях использовались такие модели, как иерархическая сеть внимания на основе трансформеров [44]. Исключительно в таких работах, как [45], изучалась Roberta-BiLSTM и использовался [46] клинический анализ для выделения признаков и классификации депрессий; в таких исследованиях, как [45], [47] и [48], традиционные модели трансформеров сочетались с сетями LSTM для выделения признаков. Хотя большинство из них использовали модели внимания на основе трансформеров, мы рекомендуем использовать модели преобразования вместо моделей трансформеров. При обработке текстовых и аудиоданных модели Reformer работают аналогично моделям transformer, но быстрее и с меньшим потреблением памяти.

Для решения текущих задач авторы предлагают модель “Multi-Stream Mood Insight Encoder (MMIE)”. Эта модель использует двухуровневый подход, объединяющий текстовые клинические функции [20] с лингвистическими функциями и потоками аудиоданных, чтобы получить всестороннее представление о настроении и эмоциональном состоянии человека, одновременно улучшая его с помощью кодировщика Reformer. Эффективное локальное самонаблюдение Reformer с механизмом локально-чувствительного хэширования (LSH) позволяет MMIE эффективно обрабатывать длинные последовательности данных, анализуя как короткие, так и развернутые высказывания на предмет тонких эмоциональных сигналов. Кроме того, MMIE включает когнитивную информацию, извлекая лингвистические шаблоны и эмоциональные выражения из текстовых данных, а также улавливая акустические особенности и просодические характеристики из аудиовходов.

Объединяя эти методы в единую структуру, MMIE стремится обеспечить целостную оценку депрессивных симптомов индивида, позволяя проводить точную и своевременную классификацию. Благодаря тщательному обучению и валидации набора данных DAIC-WOZ, MMIE стремится предложить надежный инструмент для диагностики депрессии, что в конечном итоге способствует улучшению оценки психического здоровья и стратегий вмешательства.

Цели и содержание исследования.

Цели исследования:

1) обеспечить свежий и усовершенствованный подход к классификации мультимодальных депрессий на основе текста и аудио;
2) изучить эффективность различных языковых моделей с сочетанием звуковых признаков для идентификации депрессии;
3) изучить потенциал предлагаемого метода для клинической диагностики психических расстройств.

Содержание исследования.

1) используя LLMS и модели Reformer, авторы предоставляют двухуровневые кодировки объектов для текстовой модальности; сначала авторы экспериментировали с извлечением текстовых признаков, используя четыре языковые модели: BERT [17], MentalBERT [18], MentalRoBERTa [45], PsychBERT [19] и ClinicalBERT [20]; авторы также рекомендуют вводить лингвистические характеристики в кодеры на основе Reformer; далее эти вложения затем объединяются для обнаружения депрессии на основе текста и звука с использованием модели reformer;
2) авторы исследовали наиболее продвинутые текстовые модели для диагностики депрессии и обнаружили, что большая языковая модель ClinicalBERT зарекомендовала себя лучше других моделей, как подходящий и экономичный способ обработки больших массивов данных в сочетании с аудиофильмами Reformer encoders; наконец, авторы продемонстрировали, что использование рекомендованного ими метода “Multi-Stream Mood Insight Encoder (MMIE)” превосходит современные работы по последнему слову техники; этот метод многообещающий для клинической диагностики психических заболеваний, насколько известно авторам, никто не проводил такого тщательного анализа;
3) благодаря использованию LLM ClinicalBERT в качестве первого вклада, предложенная стратегия улучшила как производительность, так и точность в задаче классификации депрессии; в довольно широком диапазоне условий модель показала себя хорошо.

Статья структурирована следующим образом: в разделе 2 анализируется существующая литература, в разделе 3 приводится методология, в разделе 4 представлены результаты, в разделе 5 – обсуждение полученных результатов, раздел 6 является заключением.

1. Анализ сопутствующих работ.

A. Обнаружение депрессии на основе текста. Ранний скрининг на депрессию имеет решающее значение для предотвращения членовредительства и самоубийств, но современные методы сталкиваются с такими проблемами, как переоснащение модели и неточная идентификация из-за соображений конфиденциальности и ограниченности ресурсов. Огромное количество текстовых данных в социальных сетях имеет значительную практическую ценность, что побудило к недавним исследованиям их потенциала для создания автоматизированных систем обнаружения депрессии. Многочисленные исследования [14, 22, 49-58], подтверждают целесообразность использования анализа текстовых сообщений для прогнозирования депрессии. В некоторых исследованиях используются специализированные наборы данных, такие, как DAIC-WOZ [59], в то время как другие анализируют текстовые данные, полученные непосредственно из платформ социальных сетей.

От традиционных трудоемких методов извлечения текстовых объектов перешли к автоматизированным подходам, таким, как обученные встраивания Word2Vec и TF-IDF (термин, обратный частоте документа) в глубокое обучение, демонстрируя эффективность при обработке многомерных данных [14, 50]. В нескольких исследованиях используются многоуровневые, вариационные и K-конкурентные автоэнкодеры (Autoencoders) для придания особого значения снижению шума, тем самым повышая надежность модели и превосходя Word2Vec в задачах регрессии и классификации текста [51-57]. Включение линейных или нелинейных скрытых функций из автоэнкодеров повышает точность классификации, в то время как сверточные автоэнкодеры в сочетании с нейронными сетями повышают производительность в таких задачах, как MNIST и распознавание объектов [49, 58]. Используя гетерогенный график и потерю фокуса для извлечения информации из диалога, некоторые модели механизма внимания, такие, как нейронная сеть с гетерогенным графиком внимания [60], пытались идентифицировать депрессию с помощью контекстуальных

сигналов, и их оценка в наборе данных DAIC-WOZ показала, что это может помочь врачам распознать депрессию в стенограммах клинических интервью.

Модели гибридных случайных лесов на основе BERT (BERT-RF) [61] недавно были использованы для разработки функций раннего обнаружения по сообщениям в социальных сетях. Эти модели также требовали высокой точности в 99%, аналогичной моделям в исследовании [14], вносящим значительный вклад в аналитику психического здоровья.

Также были обнаружены значительные препятствия в области самостоятельного обучения представлению речи, такие, как нехватка словарного запаса и непредсказуемость длины звуковых единиц во входных высказываниях. Хотя это был многообещающий шаг, метод Хьюберта [62], уникальная фаза автономной кластеризации, которая обеспечивает выровненные целевые метки для потери прогноза, подобной BERT, оказался не очень успешным в решении этих проблем. Между тем, в области автоматического выявления депрессии недавние успехи зависят от использования больших языковых моделей, таких, как BERT, MentalBERT, MentalRoBERTa, ClinicalBERT и PsychBERT, предварительно обученных на обширных наборах данных. Эти модели свидетельствуют о переходе к более тонким и эффективным подходам к пониманию и решению проблем психического здоровья [11, 17-20, 63-64].

Предыдущие исследования показали, что модели гибридной сверточной нейронной сети (CNN) и долговременной кратковременной памяти (LSTM) эффективны для анализа текстовых данных с платформ социальных сетей с целью выявления депрессии. [39, 65]. Модели CNN и LSTM были интегрированы Вани и др. [14] для классификации депрессии на основе данных, полученных из текстов в социальных сетях. Более того, другие исследования выступали за принятие двунаправленной модели LSTM (Bi-LSTM) [13, 66].

B. Обнаружение депрессии на основе мультимодальных данных. Выявление депрессии, имеющее решающее значение в современных условиях, сопровождается ростом числа исследований, посвященных мультимодальному анализу данных. Платформа интермодального слияния для обнаружения депрессии на основе корпуса, называемая IIFDD и недавно использованная [67], объединяет внутри-модальные и интер-модальные функции для обнаружения депрессии в различных наборах данных. Здесь также предпринимались такие усилия, как [68], которые объединили аудио, видео и семантические функции для создания модели регрессии по гауссовой лестнице, в то время как в [69] использовали сеть LSTM для фиксации взаимодействия аудио- и текстовых функций для обнаружения депрессии. Примечательно, что даже с добавлением стробирующих ячеек с использованием LSTM, недостатки моделей RNN, такие, как проблемы с запоминанием длинных последовательностей и захватом сложных зависимостей, остаются. Кроме того, поскольку каждый прямой проход зависит от результата обработки предыдущего временного шага, последовательная топология сети создает трудности для эффективных параллельных вычислений в сети LSTM. Поэтому, чтобы обойти эти проблемы и изучить особенности поведения у пациентов с депрессией, гибридные модели CNN-LSTM были обучены на мультимодальных данных, которые содержали как текст, так и аудиовходы [70, 71], хотя в некоторых мультимодальных подходах к диагностике депрессии явно использовались модели CNN, интегрирующие функции MFCC и спектрограммы, и проверенные с использованием наборов данных MODMA и RAVDESS. [72].

Учитывая необходимость объяснимости в области медицины, в [73] интегрировали модель самоконтроля с распадом GRU и оценили ее эффективность. Аналогичным образом, в [74] ввели интерпретируемую мультимодальную парадигму, основанную на характеристиках в анализе данных о здоровье, чтобы улучшить обобщение и диагностические данные для скрининга тревожности. Полученные результаты показали, что алгоритмы глубокого

обучения могут быть полезны для цифровых систем поддержки принятия решений, что может оказать положительное влияние на качество и стоимость здравоохранения.

Метод выявления депрессии без использования заранее установленных вопросников, таких, как вопросники о состоянии здоровья пациентов (PHQ-8/9), был представлен в [75], где предложена аддитивная кросс-модальная сеть внимания для анализа аудио- и текстовых данных, которые были сопоставлены с наборами данных DAIC-WOZ и EATD-Corpus.

C. Модели на основе трансформеров для обнаружения разряжения. Исследования по выявлению депрессии на основе мультимодальных данных имеют решающее значение из-за распространенности депрессии, а устройства Интернета медицинских вещей предлагают полезные ресурсы данных. Другие авторы использовали дополнительные наборы данных для мультимодального распознавания эмоций, такие, как системы распознавания на основе ЭЭГ; самый последний вклад был внесен в [76]. Для распознавания эмоций на ЭЭГ в [76] создали LResCapsule, решающую проблемы с ограниченными наборами данных DEAP и DREAMER и определением пространственных объектов. В то время, как в [77] предложили интегрировать аудио-, видео- и дистанционные фотоплетизмографические сигналы (rPPG, rPPG), они сделали это, используя модули на основе трансформеров для улучшения объединения и представления объектов и deep CNNs для извлечения объектов. Через пять лет после того, как Google представила модели transformer [42], их использование распространилось на несколько областей и постоянно превосходило модели на основе LSTM [11, 15, 41, 44]. По сравнению с альтернативными архитектурными моделями, модели на основе трансформеров [11, 21, 23, 36, 41] постоянно демонстрируют более высокую производительность.

Для того, чтобы моделировать разговоры и выявлять различные психические заболевания по ходу диалога, в [22] ввели иерархический классификатор глубокой нейронной сети, основанный на внимании (LSTM). Гибридная трансформаторная сеть, содержащая BERT и Bi-LSTM, а также MLP в конце, была использована Дару и др. [47] для реализации методологии контекстуального обнаружения депрессии для идентификации текстов, наводящих на мысль о депрессии. Чтобы извлечь характеристики из последовательностей депрессивных текстов, Чжан и др. [45] одновременно использовали гибридную модель глубокого обучения Robert-BiLSTM. Существует настоятельная необходимость исследовать эти новые методы, даже несмотря на продолжающиеся попытки выявления депрессии с использованием корпуса DAIC-WOZ, которые еще не привели к этому.

Хотя модели трансформеров продемонстрировали успех в развертывании обнаружения депрессий, их ограниченность в обработке длинных последовательностей побудила рассмотреть модели преобразователей [43]. Модели Reformer значительно снижают требования к памяти по сравнению с классическими преобразователями за счет эффективной аппроксимации механизмов внимания с использованием хеширования, чувствительного к локальности (LSH). В результате возможна обработка больших наборов данных без проблем с памятью. Основанный на LSH механизм внимания Reformer models помогает лучше извлекать релевантную информацию из текста и слухового восприятия, что может помочь выявить тонкие закономерности, указывающие на депрессию. Модели Reformer могут быть более устойчивы к шуму в текстовых и аудиоданных, поскольку их процесс самоконтроля менее восприимчив к шуму и нерелевантной информации. Модели Reformer, прошедшие обучение на обширных наборах аудио- и текстовых данных, могут улучшить свои показатели на наборе данных DAIC-WOZ за счет использования обучения переносу, что повысит точность идентификации депрессии.

Для решения этих проблем авторы предлагают новую модель, которая использует кодировщик Reformer [43] для эффективного захвата зависимостей на большом расстоянии. Эта работа подчеркивает серьезность депрессии и ее последствий в ответ на настоятельную

потребность в непредвзятой и надежной модели “Multi-Stream Mood Insight Encoder (MMIE)”. Чтобы решить проблемы с текстовыми расшифровками и длинными аудиозаписями для анализа, предлагаемая парадигма, поддержанная методологией, объединяет Reformer encoder. Модель значительно улучшается со средней абсолютной ошибкой (MAE) 3,421 и среднеквадратичной ошибкой (RMSE) 4,641 в наборе данных DAIC-WOZ.

2. Методология исследования

2.1 Описание. Набор данных для интервью по анализу дистресса – Wizard-of-Oz (DAIC-WOZ) [32, 59] доступен на веб-сайте Института креативных технологий Университета Южной Калифорнии. Это жизненно важный ресурс для исследований аффективных вычислений и автоматизации прогнозирования проблем, связанных с психическим здоровьем. Доступны как расширенная, так и более ранняя версии наборов данных, которые были разработаны для конкурса аудиовизуальных эмоций (AVEC-2019). Они включают физиологические данные, аннотации к голосу и выражению лица, стенограммы текстовых, видео- и аудиоинтервью с людьми, испытывающими психиатрический дистресс [59, 78]. 189 сеансов в наборе данных продолжительностью от 7 до 33 минут взяты из интервью с "Волшебником страны Оз", которые проводились в контролируемой обстановке при поддержке Элли, виртуального ассистента. Набор данных содержал 189 выборок, из которых 133 были отнесены к категории "без депрессии" и 56 – "с депрессией".

2.2 Предварительная обработка данных и извлечение признаков. На первом этапе каждый раз, когда задавался новый вопрос, виртуальный ассистент Элли автоматически извлекала пары “вопрос-ответ” и связанные с ними временные метки из необработанных стенограмм. Эти временные метки облегчили разделение необработанных стенограмм на группы, каждая из которых представляла собой пару вопросов и ответов. После того, как были выделены пары “вопрос-ответ”, мы использовали обработку естественного языка (NLP) для извлечения атрибутов участников. Предварительная обработка текстовых данных включала лингвистические методы, в т.ч. удаление специальных символов и стоп-слов в дополнение к преобразованию текста в нижний регистр [79]. Для подготовки данных для моделей NLP были использованы методы, включающие токенизацию, лемматизацию и стемминг. Чтобы максимизировать производительность модели, удалению стоп-слова был отдан приоритет по эффективности, за которым следовали стемминг и лемматизация. Устранение неоднозначности значения слова (WSD) и разбиение на фрагменты касались структуры предложения и двусмысленности, в то время как пометка частей речи давала синтаксическую информацию. Методы синтаксического анализа выполнялись такими модулями Python, как NLTK, spaCy, TextBlob и hero [35, 80]. Намеренное оставление вопросов без ответов во время обсуждения использовалось для сохранения контекстуального обучения и уважения автономии участников при обсуждении деликатных тем. Количество абсолютистских слов и местоимений от первого лица было получено с помощью инструментария Python NLTK toolkit. Блок-схема, иллюстрирующая предлагаемые этапы предварительной обработки данных, изображена на рисунке 1.

BERT (представления двунаправленных кодеров от Transformers), разработанные Google AI, являются одной из наиболее известных языковых моделей [17]. Теперь мы можем использовать предварительно обученные языковые модели для приложений в области психического здоровья, такие, как MentalBERT, MentalRoBERTa [18], PsychBERT [19] и ClinicalBERT [20].

Эти модели были обучены с использованием большого набора данных сообщений Reddit, связанных с психическим здоровьем. Они построены на архитектуре BERT masked language model (MLM). В этой работе авторы извлекали текстовые объекты с помощью MentalBERT, MentalRoBERTa, PsychBERT и ClinicalBERT на платформе Hugging Face. Из-за

разного объема стенограмм извлеченные атрибуты каждого участника имели разную длину. После выбора всей пары “вопрос-ответ” вместе с их временными метками авторы предварительно тщательно обработали текст в рамках предложенной методологии, выбрав каждую пару “вопрос-ответ” вместе с отметкой времени. После этого этапа авторы использовали BERT, ClinicalBERT, MentalBERT, MentalRoBERTa и PsychBERT для извлечения признаков. Кроме того, они были объединены с извлеченными языковыми элементами, такими, как местоимения от первого лица и абсолютное количество слов.

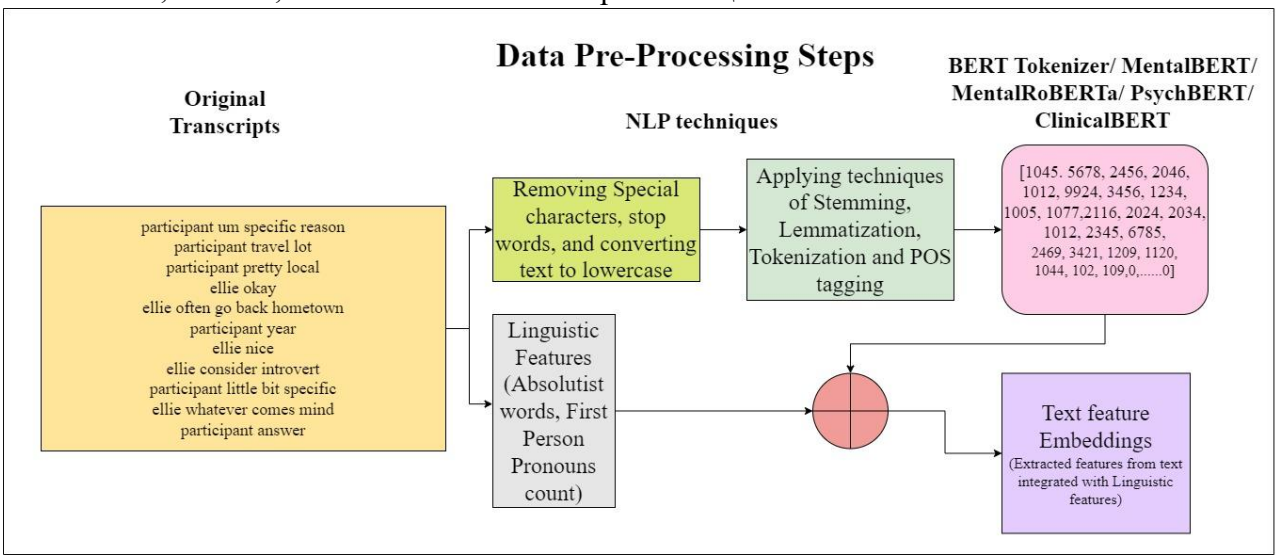

Рис. 1. Этапы предварительной обработки данных

Набор данных содержал предварительно обработанные звуковые характеристики, такие, как COVRAEP / FORMANT характеристики. (COVRAEP / FORMANT – э*то методы извлечения признаков.*) Эти конкретные наборы звуковых функций были дополнительно объединены с извлеченными текстовыми функциями и лингвистическими характеристиками, чтобы служить окончательным тензором для ввода в предлагаемую модель.

2.3 Кодировщик модели Reformer. Математическая формулировка кодировщика модели Reformer такова:

1) вложения входных данных: пусть X представляет входную последовательность, состоящую из N токенов, каждый токен представлен $d_{модели}$. Таким образом, X представляет собой матрицу формы (N, $d_{модель}$);
2) позиционные вложения: аналогично оригинальному transformer, позиционные кодировки добавляются к входным вложениям для предоставления позиционной информации; пусть P представляет матрицу позиционных кодировок формы (N, $d_{модели}$), где каждая строка представляет позиционную кодировку для токена;
3) вход кодирующего устройства: вход кодирующего устройства Reformer представляет собой сумму входных вложений X и позиционных кодировок: P: $X_{вход}X + P$;
4) обратимые слои: в отличие от стандартного преобразователя, Преобразователь использует обратимые слои для обеспечения эффективной обработки длинных последовательностей с использованием памяти; обратимый слой состоит из двух подуровней: Прямой подуровень (F); Обратимые слои; Этот слой применяет серию операций к входным данным без изменения их размера; обратимая остаточная сеть (RRN): этот уровень обратимо применяет последовательность преобразований к входным данным;
5) локальный самоконтроль с локально-чувствительным хэшированием (LSH): вместо стандартного механизма самоконтроля, используемого в transformer, Преобразователь использует LSH для приближения полного самоконтроля более экономичным способом

с использованием памяти; LSH хэширует входные векторы в сегменты и обрабатывает данные только в пределах соседних сегментов, снижая вычислительную сложность;

6) сеть прямой связи (FFN): каждый обратимый уровень обычно включает в себя нейронную сеть прямой связи, которая применяет нелинейное преобразование к каждому токену независимо;
7) нормализация слоя и остаточные соединения: нормализация слоя и остаточные соединения применяются после каждого обратимого слоя для стабилизации обучения и облегчения градиентного потока.

Математически преобразователь кода можно представить в виде серии обратимых уровней:

$$X_{\text{вход}} = X + P \quad (1)$$

$$X_{\text{выход}} = \text{преобразователь преобразователя } (X_{\text{вход}}) \quad (2)$$

$$X_{\text{выход}} = F(RRN(X_{\text{вход}}) \quad (3)$$

где $X_{\text{выход}}$ представляет выходной сигнал преобразователя, а F и RRN представляют операции, выполняемые в пределах каждого обратимого уровня.

2.4 Предлагаемая MMIE модель. Авторы предлагают новый подход к персонализированному выявлению депрессии с использованием аудио- и текстовых модальностей: гибридную модель глубокого обучения, основанную на клинической модели Reformer, настроенной на основе Alberta. Она воплощает в себе архитектуру нейронной сети, которая использует возможности сетей Reformer наряду с обширными языковыми методологиями извлечения признаков. Значимые текстовые элементы были извлечены с помощью LLMS и объединены вместе с лингвистическими и звуковыми элементами перед подачей в качестве входных данных в модель Reformer. Авторы обнаружили, что клинический анализ был наиболее успешным методом. Текстовые функции, полученные из больших языковых моделей, лингвистические функции и аудиоданные были обработаны с помощью Reformer encoder для создания кодировок на основе reformer. Технологическая схема предлагаемой модели изображена на рисунке 2.

2.5 Алгоритм.

Ввод:

Представления функций:

X^{i}_{BERT} извлечен из варианта i BERT (где i означает BERT, MentalBERT, MentalRoBERTa, PsychBERT ИЛИ ClinicalBERT).

$X_{аудио}$, извлеченное из аудиоданных (COVAREP и ФОРМАНТ).

$X_{лингвистический}$ для лингвистического признака.

Изучаемые веса:

X^{i}_{BERT}, $X_{аудио}$ и $X_{лингвистический}$ для каждого типа объектов.

Объединение:

$$X_{\text{concat}} = (X^{1}_{BERT,}X^{2}_{MentalBERT,}\dots \text{ИЛИ} \dots X^{5}_{ClinicalBERT,}) + X_{\text{аудио}} + X_{\text{лингвистика}}$$

Reformer-based LSH Attention:

$$F_{Reformer} = \text{ReformerLSHAttention}(X_{\text{объединение}})$$

Результат:

Классификация депрессии:

$$\hat{y} = \text{Классификатор}(F_{Reformer})$$

где классификатор представляет собой классификатор, обученный с использованием $F_{Reformer}$ в качестве входных признаков для классификации депрессий,

y = классификационная метка депрессии;

$\hat{y}$ = метка классификации прогнозируемой депрессии.

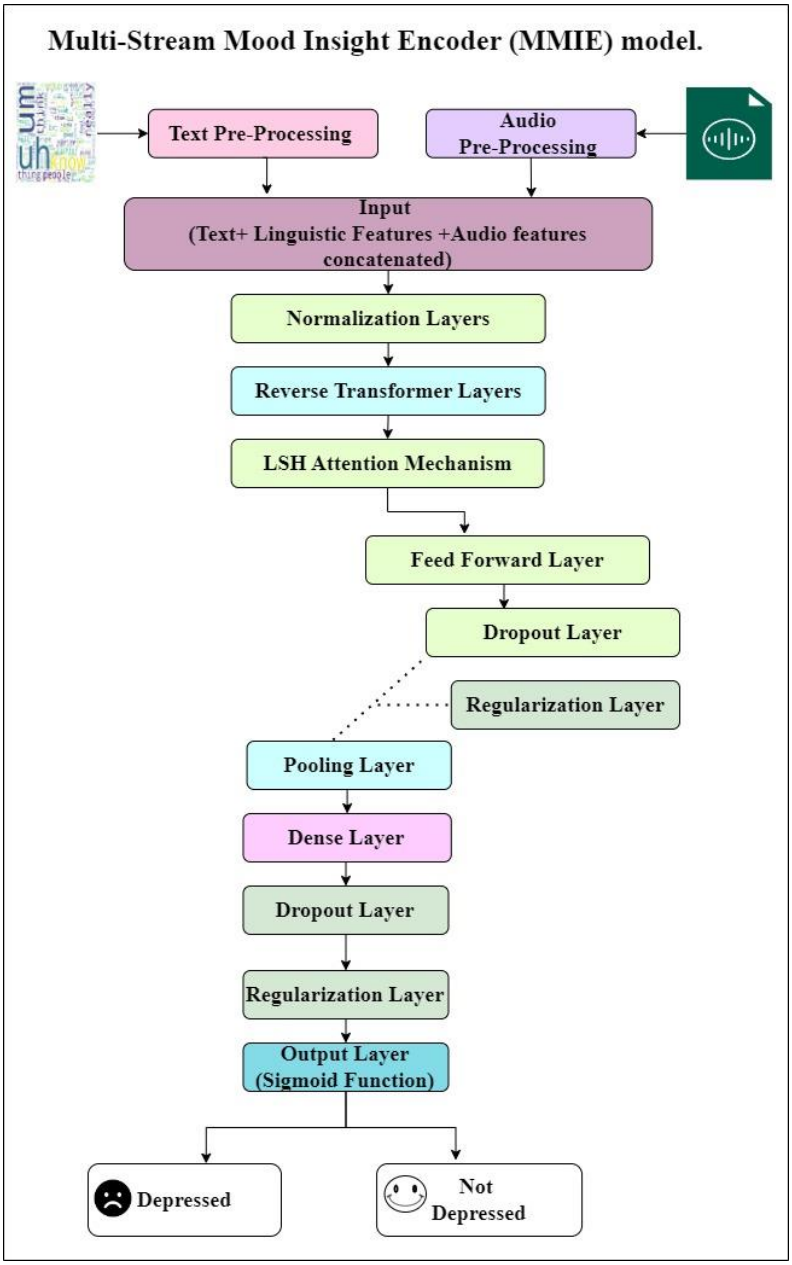

Рис. 2. Блок-схема модели обратного трансформера: многопоточная модель Mood Insight Encoder (MMIE).

2.6 Экспериментальные показатели обучения и оценки. Предложенная модель обучалась на графическом процессоре, что повысило вычислительную эффективность. Процесс обучения включал 200 итеративных периодов обучения. Таблица 1 иллюстрирует экспериментальную установку, использованную в нашем исследовании.

Таблица 1. Экспериментальная установка

Используемое аппаратное и программное обеспечение	**Технические характеристики**
GPU	NVIDIA GeForce RTX 2080 SUPER
RAM	128 ГБ
OS	Windows 11/ Linux
Язык программирования	Python

Авторы оптимизировали производительность модели, скорректировав гиперпараметры с использованием методов поиска по сетке, и использовали RMSProp optimizer для лучшей

оптимизации модели; чтобы обеспечить обобщаемость и предотвратить переобучение, также использовалась процедура ранней остановки регуляризации. Распространение среднеквадратичного значения, или оптимизатор RMSProp, и алгоритм градиентного спуска с импульсом были сопоставимы. Колебания в вертикальной плоскости были ограничены оптимизатором RMSProp. Поскольку это может ускорить обучение, предложенный авторами алгоритм может выполнять большие горизонтальные шаги и быстрее достигать сходимости.

Градиентный спуск отличался от RMSProp способом расчета градиентов. Каждый вес в сети изучался с разной скоростью в результате того, что градиент нормировался на знаменатель алгоритма обновления веса. Это может смягчить проблемы, связанные с другими стратегиями оптимизации на основе градиента, такими, как исчезновение или разрыв градиентов.

Политика обновления RMSProp определяется:

$$S_{dw} = \beta S_{dw} + (1-\beta)dw^2 \quad (4)$$

$$W = W - \alpha \frac{dw}{\sqrt{S_{dw}}} \quad (5)$$

где:

S_{dw} = скользящее среднее квадрата градиента.

β = коэффициент забывания.

d_w = градиент функции потерь, зависящей от веса.

α = скорость обучения.

W = параметр веса.

Для оценки результатов использовались три показателя точности – F1, Precision и Recall. Более высокий показатель точности означал улучшение характеристик модели. Поведение регрессионной модели было дополнительно исследовано с использованием результатов средней абсолютной ошибки (MAE) и среднеквадратичной ошибки (RMSE).

$$\text{Точность (A)} = \frac{TP+TN}{TP+TN+FP+FN} \quad (6)$$

$$\text{Точность (P)} = \frac{TP}{TP+FP} \quad (7)$$

$$\text{Вспомнить (R)} = \frac{TP}{TP+FN} \quad (8)$$

$$\text{Оценка F1} = \frac{2*\text{Точность}*\text{Припоминание}}{\text{точности}+\text{Припоминание}} \text{ или } \frac{2*TP}{2*TP+FP+FN} \quad (9)$$

где TP = истинно положительный результат, TN= истинно отрицательный результат, FP = ложноположительный результат и FN = ложноотрицательный

$$\text{Средняя абсолютная погрешность (MAE): } = \frac{1}{n}\sum_{i\cdot=1}^{n}(y_i - \hat{y}_i) \quad (10)$$

$$\text{среднеквадратичная ошибка (MSE): } = \frac{1}{n} * \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (11)$$

где n = количество наблюдений в наборе данных.

y_i= фактическое значение i-го наблюдения.

$\hat{y}_i$= прогнозируемое значение i-го наблюдения.

$$\text{Среднеквадратичная ошибка (RMSE): =sqrt (MSE)} \quad (12)$$

3. Результаты. Авторы провели серию экспериментов по объединению текстовых функций на основе BERT со звуковыми функциями, чтобы оценить их эффективность и воспользоваться возможностями, встроенными в кодировщики Reformer с глубоким обучением (DL). Эффективность предлагаемого метода систематически оценивалась с использованием набора данных DAIC_WOZ, состоящего из двух наборов по 10 экспериментов в каждом. Испытания, которые соответствовали элементам из BERT, MentalBERT, MentalRoBERTa, PsychBERT и ClinicalBERT, были названы моделями, основанными на реформерах (Reformers). Подробное описание результатов этих исследований приведено в таблицах 2, 3.

Таблица 2. Результаты тестирования для LLM и классификации звуковых функций для обнаружения депрессии в DAIC-WOZDAIC-WOZ

LLM	**F1**	**Precision**	**Recall**	**MAE**	**RMSE**
BERT	0,8471	0,8410	0,8614	6,84	10,86
BERT + Аудио (COVAREP + ФОРМАНТ)	0,8615	0,8511	0,8781	6,42	9,97
MentalBERT	0,8575	0,8556	0,8614	5,98	7,98
MentalBERT + Аудио (COVAREP + FORMANT)	0.8962	0.8925	0.9021	5.23	7.21
MentalRoBERTa	0.8676	0.8501	0.8896	4.26	7.26
MentalRoBERTa + Аудио (COVAREP + FORMANT)	0.9054	0.9001	0.9105	3.68	6.06
PsychBERT	0.9066	0.9025	0.9101	6.71	9.52
PsychBERT+ Аудио (COVAREP + FORMANT)	0.9144	0.9110	0.9214	5.44	7.32
ClinicalBERT	0.9288	0.9310	0.9296	3.68	5.01
ClinicalBERT+ Аудио (COVAREP + FORMANT)	**0.9538**	0.9499	0.9574	**3.42**	**4.64**

Таблица 3. Гиперпараметрическая оптимизация предлагаемой модели

MentalBERT, MentalRoBERTa, PsychBERT, ClinicalBERT +аудио характеристики (COVAREP + FORMANT)	
Layer Name	**Parameter Settings**
Input Size (Text +Audio)	786+ 1684
Optimizer	RMSprop
Learning Rate	0.001
Batch Size	256
Epochs	200
Regularization (L2)	0.001
Dropout Rate	0.2
Number of Dense Layers	4
Number of Reverse Transformer Layers	2
Number of Locality-Sensitive Hashing (LSH) Attention	2
Units in Dense Layers	512, 16, 256, 256
Activation Function in Dense Layers	ReLU

Units in Output Layer	1
Activation Function in Output Layer	Sigmoid
Early Stopping	Yes
Early Stopping Monitor	Validation Loss
Early Stopping Patience	3

5. Обсуждение. Авторы оценили производительность предложенной ими модели с помощью набора для разработки DAIC-WOZ. Результаты, как показано в таблицах 2-5, показывают, насколько хорошо работала модель по сравнению с базовым уровнем и другими подходами. Прежде всего, предложенная авторами модель превзошла методы, предложенные в более ранних исследованиях. Сравнивая его с базовым экспериментом против Ореши и др. [12] и семью тщательно подобранными методологиями из литературы – методами, которые были выбраны за их лучшую производительность или совместимость с нашим экспериментальным подходом, – авторы эмпирически сравнили предложенную модель с ними в наборе данных DAIC. Результаты других моделей взяты из соответствующих оригинальных исследований:

1) Ореши и др. [12] предложили многозадачную модель обучения, которая использует общий уровень LSTM для отличия депрессии от мультимодальных данных;
2) Фан Х. и др. [77] извлекли необходимые высокоуровневые видео- и аудиофункции с использованием глубоких сверточных нейронных сетей (CNN), и окончательные задачи обнаружения депрессии были реализованы путем подачи представлений функций всех модальностей в сеть многослойного персептрона (MLP);
3) Лин и др. [81] рекомендовали использовать двунаправленную сеть долговременной кратковременной памяти (Bi-LSTM) со слоем внимания для обработки лингвистического контента, полностью подключенную сеть, интегрирующую выходные данные предыдущих двух моделей для оценки депрессивного состояния, и одномерную сверточную нейронную сеть (1D-CNN) для обработки речевых сигналов;
4) Сан Г. и др. [82] предложили неконтролируемый автоэнкодер на основе трансформаторов для разработки встраиваний на уровне предложений из аудиовизуальных функций на уровне кадра; они также предложили сеть глубокого объединения функций, использующую кросс-модальный преобразователь для интеграции текстовых, аудио- и видеофункций;
5) Ийортсуун, Н. и др. [75] использовали как аудио, так и текстовые данные, предложив аддитивную сеть кросс-модального внимания для изучения и подбора соответствующих весов, которые наилучшим образом отражают кросс-модальные взаимодействия и взаимосвязи между обеими функциями, используя Bi-LSTM в качестве основы обеих модальностей;
6) Ли, М. и др. [60] создали гетерогенный граф (DSE-HGAT), используя исключительно текстовую модальность, моделируя депрессивное состояние участника и объединяя фрагменты депрессивных подсказок с использованием сети graph attention network;
7) Илиас Л. и др. [11] предложили метод, в котором использовался вентиль мультимодальной адаптации для создания комбинированных вложений, которые затем загружаются в модель BERT (или MentalBERT);
8) Дас А. и др. [72] применили мультимодальный подход объединив функции MFCC, а также функции спектрограммы, извлеченные из аудиофайла, с помощью новой сети CNN;

9) J. Ye. и др. [84] – аудиовизуальная мамба с прогрессивным слиянием для выявления мультимодальной депрессии, получившая название DepMamba;
10) результаты классификации обнаружения разрежения с использованием различных LLM, звуковых функций и эффективного преобразователя с обратным преобразованием, называемого Reformer, показаны в таблице 2; гиперпараметры модели Реформера, которые были определены для получения этих результатов, перечислены в таблице 3. Авторы оптимизировали производительность модели, скорректировав гиперпараметры с использованием подходов поиска по сетке.

Таблица 4. Сравнение с предыдущими работами по выявлению депрессии как проблемы регрессии

Существующие модели, подходы,	**Применяемые методы**	**MAE**	**RMSE**
Исходные данные Оуреши, С. А. и др. [12]	Мультимодальны общий слой LSTM фьюжн	3.49	4.71
Фан, Х. и др. [77]	CNN+ трансформатор на основе модели	7.05	9.45
Лин и др. [81]	аудио + текст, Bi-LSTM модель	3.88	5.44
Сан, Г. и др. [82]	мультимодальны DDFN	3.78	5.35
Ийортсуун, Н. и др. [75]	Additive Cross-Modal Attention Network (ACMA)	4.65	/
Наша модель только с текстом и реформаторскими моделями	ClinicalBERT + Reformer Encoder и классификатором	3.684	5.019
Наша модель с текстовой и аудиомодальностью и классификатором Reformer Encoder	ClinicalBERT +Reformer Encoder и классификатором	**3.421**	**4.641**

Таблица 5. Сравнение с предыдущими работами по выявлению депрессии как classification problem

Существующие модель-ные подходы	**Применяемые методы**	**F1**	**Точность**	**Вспомнить**
Лин и др. [81]	Аудио + текст, модель Bi-LSTM	0.83	0.83	0.83
Сан, Г. и др. [82]	Мультимодальный, DDFN	0.89	0.91	.
Ийортсуун, Н. и др. [75]	Additive Cross-Modal Attention Network (ACMA)	0,82	0,79	0,86
Ли, М. и др. [60]	Текст, DSE-HGAT	0.79	0.79	0,80
Илиас, Л. и др. [11]	M-MentalBERT (top2vec)	93,06	96,12	90,18
Дас, А. и и др.. [72]	CNN	0,89	0,93	0,85
Предлагаемая модель с текстом единственный и Модели Реформер	ClinicalBERT + Reformer Encoder and Classifier	.	.	

Предлагаемая модель с текстовой и аудиомодальной модальностью и классификатором Reformer Encoder	**ClinicalBERT +Reformer Encoder and Classifier**	**0.95**	**0.94**	**0.95**

Исследование абляции. Чтобы подтвердить эффективность каждого модуля в модели авторов, проведены эксперименты по абляции с различными комбинациями модулей:

– Когда звуковые функции не доступны в модели Reformer, текстовые функции извлекаются непосредственно из LLMS и затем загружаются в модель Reformer. Использование функций аудио в сочетании с кодировщиком Reformer, использующим локально-зависимое хэширование (LSH), значительно улучшило MAE с 6,84 до 6,42 и RMSE с 10,86 до 9,97 в случае модели BERT только до модели BERT + Audio (COVAREP + FORMANT). Другие модели с различными LLMSS также показали аналогичные результаты; например, модель MentalRoBERTa с аудиокомпонентами существенно превзошла другие модели в исследованиях. Наилучшие результаты были получены при значениях MAE, которые улучшились с 3,68 до 3,42, и значениях RMSE с 5,01 до 4,64 при использовании модуля ClinicalBERT.

– Это еще раз подчеркивает преимущества мультимодальных признаков для классификации депрессии. Классификатор Reformer с текстом и звуком (COVAREP + FORMANT) неизменно превосходил текстовые функции, демонстрируя эффективность этой комбинации. Применяя эту тактику, ClinicalBERT добился максимальной производительности по всем направлениям. Результаты эксперимента еще раз подтверждают успех усилий авторов по кодированию объектов, подчеркивая совокупную эффективность кодировщиков Reformer в LLM для исследования корпусов, а также звуковых и лингвистических характеристик. Однако модели Reformer, возможно, по своей природе не способны улавливать долгосрочные временные корреляции, особенно при работе с текстовыми и аудиоданными модальностями. При объединении и использовании данных, как из текстовых, так и из аудиальных источников модели Reformer могут столкнуться с трудностями. Чтобы обойти требование объединения объектов во время обучения модели, авторы объединили объекты с помощью простой конкатенации.

– В таблице 3 приведены результаты гиперпараметрической оптимизации, в то время как в таблицах 4 и 5 представлено сравнение производительности модели. Оптимальные конфигурации гиперпараметров, определяемые с точки зрения максимальной точности классификации, включают настройку модуля LSH attention с двумя модулями и установку размера пакета на 256. Эти результаты свидетельствуют о том, что механизм внимания LSH способствует повышению точности классификации модели. Кроме того, наблюдаемые улучшения производительности подразумевают, что модель способна эффективно обрабатывать большие объемы данных и захватывать сложные шаблоны с использованием аппаратного обеспечения GPU с механизмом внимания LSH.

Контраст по отношению друг к другу. Подробное сравнение всех моделей, использованных для решения задачи, можно найти в таблицах 4 и 5. Базовая модель продемонстрировала довольно низкие показатели производительности при использовании метода ансамбля. С точки зрения MAE и RMSE модель авторов показала себя немного лучше, чем Оуреши и др. [12]. Более того, по показателям MAE, RMSE, точности, отзыва и F1 модель авторов превзошла Лин и др. [81]. Она также превосходит Сан, Г. и др. [82] и Ли, М. и др. [60] с точки зрения метрики F1, точности и отзыва. Несмотря на достижение впечатляющих

показателей текстовой модальности и очень высокой точности, Илиасу и др. [11] не хватало информации о MAE и RMSE, которые важны для оценки регрессионных задач, таких, как прогнозирование тяжести депрессии. Хотя они также получили очень высокую точность, Дас, А. и др. [72] не соответствовали требованиям MAE и RMSE модели авторов. Удивительно, но большинство ранее раскрытых базовых методологий, за исключением нескольких недавних работ, таких, как Илиас и др., 2023 [11], не содержали лингвистических функций для текстовых моделей или обнаружения мультимодальных депрессий. Чтобы обеспечить полное представление и подчеркнуть недостаточное использование кодеров Reformer, модель авторов сочетает в себе функции кодированных представлений, а именно абсолютистские слова. Интеграция лингвистических функций, возможно, также способствовала повышению производительности нашей модели. Авторы перешли с оптимизатора Adam на RMSProp (среднеквадратичное распространение), чтобы более эффективно обновлять параметры модели во время обучения.

Обобщение модели. Авторы оценили свою модель с использованием LIME (local interpretable model-agnostic explanations), чтобы получить более надежную оценку ее эффективности. LIME – это метод для прояснения предсказаний любой модели "черного ящика", включая модели Реформер. Несмотря на то, что существует компромисс между эффективностью и прозрачностью моделей машинного обучения, попытки уточнить параметры обученной модели или веса, связанные с особенностями, тем не менее, могут способствовать лучшему пониманию моделей. Многие исследователи использовали LIME, потому что он позволяет получать данные, которые легко и быстро понять [83]. На рисунке 3 показано объяснение одного из модулей модели для предсказаний, сделанных авторской моделью с использованием LIME. Отлаженный на основе крупномасштабных клинических заметок, ClinicalBERT использует контекстуальные вложения, адаптированные к предметной области, для надежного обобщения различных клинических описаний тяжести депрессии. Объединяя текстовые вложения ClinicalBERT с низкоуровневыми акустическими характеристиками (например, показателями COVRAEP / FORMANT) в мультимодальной архитектуре на основе LSTM, модель фиксирует дополнительные биомаркеры лингвистической и просодической депрессии, улучшая обобщение в зависимости от условий записи и говорящих.

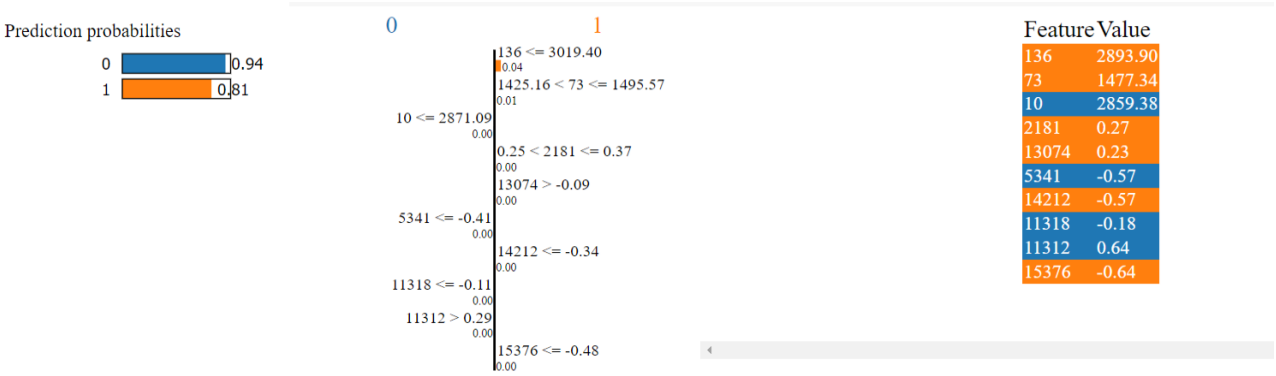

Рис. 3. Этапы предварительной обработки данных

Отчет о десятикратной категоризации, созданный с использованием предложенной методологии, также представлен в таблице 6.

Таблица 6. Сравнение с предыдущими работами по выявлению депрессии как проблеме классификации

Сложите	Государственной	Точности	Вспомнить	Ф1
1	В Депрессию	0.89	0.97	0.93

	Номера-Депрессия	0.93	0.91	0.92
2	Депрессию	0.91	0.94	0.92
	Номера-Депрессия	0.98	0.94	0.96
3	Депрессию	0.96	0.97	0.97
	Номера-Депрессия	0.98	0.96	0.97
4	Депрессию	0.95	0.96	0.95
	НомераДепрессия	0.94	0.98	0.96
5	Депрессию	0.90	0.92	0.91
	Номера-Депрессия	0.97	0.92	0.95
6	Депрессии	0.97	0.96	0.96
	Номера-Депрессия	0.98	0.95	0.96
7	Депрессии	0.95	0.96	0.95
	Номера-Депрессия	0.89	0.97	0.93
8	Депрессии	0.97	0.91	0.94
	НомераДепрессия	0.91	0.98	0.94
9	Депрессии	0.99	0.98	0.99
	Номера-Депрессия	0.91	1.00	0.95
10	Депрессия	0.96	0.94	0.95
	Номера-Депрессия	0.96	0.91	0.94
Означает	Депрессию	0.94	0.95	0.95
	Номера-Депрессия	0.94	0.94	0.95

Анализ соглашений. Анализ соответствия авторской модели дал точность обучения 94,75% и точность тестирования 95,32%. Авторы продемонстрировали примечательные результаты, оцениваемые по всем базовым показателям на указанных подтверждающих наборах, что дало оценку F1, точности и отзыва 0,9538, 0,9499 и 0,9574 соответственно. Действительно пониженные метки и прогнозы модели показывают значительную степень согласия, на что указывает значение каппа 0,717 и значение p менее 0,01. Кроме того, модель авторов улучшила показатели оценки MAE на 0,069 по сравнению с базовой моделью Оуреши и др. [12], которая использовала модели transformer, в то время как авторы использовали BERT и specific LLM for mental health, которые также являются моделями на основе transformer. Авторы считают, что повышение производительности произошло из-за использования промежуточного уровня кодировщика Reformer, который интегрировал синтаксические, семантические данные и сведения о говорящих, извлеченные из уровня LLM, и объединил их с лингвистическими функциями. Эти важные выводы имеют решающее значение для понимания их роли в анализе и классификации текстов. В частности, в сочетании с ClinicalBERT они превосходно помогают извлекать контекстно-зависимые функции из текста, еще повышая их эффективность.

Заключение. Тяжелое депрессивное расстройство (МДД) в настоящее время поражает значительную часть населения земного шара. В этой статье анализируется и предлагается уникальная методология – модель “Multi-stream mood insight encoder (MMIE)”, которая потенциально позволяет добиться значительных успехов в выявлении мультимодальной депрессии. Рекомендуемый подход MMIE является эффективным методом, который имеет перспективы для клинического выявления проблем с психическим здоровьем. Это первое исследование, в котором применяется стратегия кодирования функций, сочетающая звуковые

функции на основе COVRAEP и FORMANT с такими функциями, как absolutist words для полного представления, и предварительно обученный LLM под названием ClinicalBERT. Специальный кодировщик Reformer, который разработали авторы, используется в исследовании для классификации депрессии. Используя этот новый подход, авторы надеются внести значительный вклад в решение проблемы выявления депрессии, предоставив точку зрения для понимания и лечения этой распространенной проблемы психического здоровья.

Авторская модель, основанная на Реформере, потенциально обладает лучшей обобщаемостью, из-за ее способности фиксировать долгосрочные зависимости в текстовых и аудиоданных, что может иметь решающее значение для выявления тонких лингвистических маркеров депрессии. Включение аудиоданных (ClinicalBERT + кодировщик Reformer и классификатор с текстовой и аудио-модальной модальностью) в авторскую модель привело к значительному улучшению всех показателей по сравнению с текстовой версией. Это говорит о том, что модель эффективно извлекла уроки из обеих модальностей, потенциально приводя к более полному пониманию депрессии. Однако одним из недостатков является то, что количество и качество используемых обучающих данных может повлиять на производительность всех моделей, включая авторскую. Для более надежного сравнения необходимы исследования с использованием более крупных и разнообразных наборов данных на языках, отличных от англоязычных корпусов.

Предложенная авторами многопоточная модель “multi-stream mood insight encoder (MMIE)” продемонстрировала многообещающую эффективность для классификации депрессии, особенно в сочетании с клиническими даннымиТ и аудиоданными. Они преуспели в извлечении контекстно-зависимых функций из текстовых и звуковых модальностей, что еще больше повысило их эффективность. Это обеспечило полезный подход, который может найти применение в клинических ситуациях. Это исследование имеет важное значение в постоянных усилиях по улучшению диагностики и терапии психического здоровья, о чем свидетельствуют продемонстрированные улучшения точности и надежной работы.

Благодарности. Авторы выражают благодарность лицензии DAIC_WOZ dataset, которая значительно облегчила исследовательские усилия. Кроме того, авторы признательны библиотеке Томского государственного университета за то, что она предоставила авторам доступ к широкому спектру ценных ресурсов. «Исследование выполнено при поддержке гранта Правительства Российской Федерации (Соглашение № 075-15-2025-009 от 28 февраля 2025 г.)».

Список источников

1. Institute of Health Metrics and Evaluation. Global health data exchange (GHDx). 2021.
2. Twenge J.M., Cooper A.B., Joiner T.E., et al. Age, period, and cohort trends in mood disorder indicators and suicide-related outcomes in a nationally representative dataset, 2005–2017. Journal of abnormal psychology, 2019, vol. 128, no. 3, pp. 185–199, DOI: 10.1037/abn0000410.
3. Diagnostic and statistical manual of mental disorders. 5th ed. American psychiatric association, 2013.
4. Nestler E.J., Barrot M., DiLeone R.J., et al. Neurobiology of Depression. Neuron, 2002, vol. 34, no. 1, pp. 13–25, DOI: 10.1016/S0896-6273(02)00653-0.
5. Cohen K. Absolutist thinking and depression, 2019.
6. Blazer D.G. Psychiatry and the oldest old. American journal of psychiatry, 2000, vol. 157, no. 12, pp. 1915–1924, DOI: 10.1176/appi.ajp.157.12.1915.
7. Zarate C.A., et al. A Randomized Trial of an N-methyl-D-aspartate Antagonist in treatment-resistant major depression. Archives of general psychiatry, 2006, vol. 63, no. 8, p. 856, DOI: 10.1001/archpsyc.63.8.856.
8. Chattopadhyay S. A neuro-fuzzy approach for the diagnosis of depression. Applied Computing and Informatics, 2017, vol. 13, no. 1, pp. 10–18, DOI: 10.1016/j.aci.2014.01.001.
9. Joshi M.L., Kanoongo N. Depression detection using emotional artificial intelligence and machine learning: A closer review. Materials Today: Proceedings, 2022, vol. 58, pp. 217–226, DOI: 10.1016/j.matpr.2022.01.467.
10. Miao X., Li Y., Wen M., Liu Y., et al. Fusing features of speech for depression classification based on higher-order spectral analysis. Speech Communication, 2022, vol. 143, pp. 46–56, DOI: 10.1016/j.specom.2022.07.006.

11. Ilias L., Mouzakitis S., Askounis D. Calibration of transformer-based models for identifying stress and depression in social media. IEEE Transactions on computational social systems, 2023, pp. 1–12, DOI: 10.1109/TCSS.2023.3283009.
12. Oureshi S.A., Dias G., Saha S., Hasanuzzaman M. Gender-aware estimation of depression severity level in a multimodal setting. 2021 International joint conference on neural networks (IJCNN), IEEE, 2021, pp. 1–8, DOI: 10.1109/IJCNN52387.2021.9534330.
13. Firoz N., Beresteneva O.G., Vladimirovich A.S., et al. Automated text-based depression detection using hybrid ConvLSTM and Bi-LSTM model. 2023 Third International conference on artificial intelligence and smart energy (ICAIS). IEEE, 2023, pp. 734–740, DOI: 10.1109/ICAIS56108.2023.10073683.
14. Ahmad Wani M., ELaffendi M.A., Shakil K.A., et al. Depression screening in humans with AI and deep learning techniques. IEEE Transactions on computational social systems, 2023, vol. 10, no. 4, pp. 2074–2089, DOI: 10.1109/TCSS.2022.3200213.
15. Mazumdar H., Chakraborty C., Sathvik M., et al. GPTFX: A Novel GPT-3 based framework for mental health detection and explanations. IEEE Journal of biomedical and health informatics, 2023, pp. 1–8, DOI: 10.1109/JBHI.2023.3328350.
16. Meng Q., Catchpoole D., Skillicom D., et al. Relational autoencoder for feature extraction. 2017 International Joint conference on neural networks (IJCNN), IEEE, 2017, pp. 364–371, DOI: 10.1109/IJCNN.2017.7965877.
17. Devlin J., Chang M.-W., Lee K., Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding, 2018, DOI: 10.48550/arXiv.1810.04805.
18. Ji S., Zhang T., Ansari L., Fu J., et al. MentalBERT: publicly available pretrained language models for mental healthcare. Proceedings of the language resources and evaluation conference (LREC), 2022, DOI: 10.48550/arXiv.2110.15621.
19. Vajre V., Naylor M., Kamath U., et al. PsychBERT: a mental health language model for social media mental health behavioral analysis. 2021 IEEE International conference on bioinformatics and biomedicine (BIBM), 2021, pp. 1077–1082, DOI: 10.1109/BIBM52615.2021.9669469.
20. Huang K., Altosaar J., Ranganath R. ClinicalBERT: modeling clinical notes and predicting hospital readmission, 2019.
21. Saha T., Ramesh Jayashree S., Saha S., et al. BERT-caps: a transformer-based capsule network for tweet act classification. IEEE Transactions on computational social systems, 2020, vol. 7, no. 5, pp. 1168–1179, DOI: 10.1109/TCSS.2020.3014128.
22. Saha T., Reddy S.M., Saha S., et al. Mental health disorder identification from motivational conversations. IEEE Transactions on computational social systems, 2023, vol. 10, no. 3, pp. 1130–1139, DOI: 10.1109/TCSS.2022.3143763.
23. Li M. , et al. TrOCR: Transformer-based optical character recognition with pre-trained models. Proceedings of the AAAI conference on Artificial Intelligence, 2023, vol. 37, no. 11, pp. 13094–13102, DOI: 10.1609/aaai.v37i11.26538.
24. Anindyaputri N.A., Girsang A.S. A Comparative study of deep learning models for detecting depressive disorder in tweets. Journal of system and management sciences, 2024, vol. 14, no. 3, DOI: 10.33168/JSMS.2024.0318.
25. Flint A.J., Black S.E., Campbell-Taylor I., et al. Abnormal speech articulation, psychomotor retardation, and subcortical dysfunction in major depression. Journal of psychiatric research, 1993, vol. 27, no. 3, pp. 309–319, DOI: 10.1016/0022-3956(93)90041-Y.
26. Korszun A. Facial pain, depression and stress – connections and directions. Journal of oral pathology & medicine, 2002, vol. 31, no. 10, pp. 615–619, DOI: 10.1034/j.1600-0714.2002.00091.x.
27. He L., Cao C. Automated depression analysis using convolutional neural networks from speech. Journal of biomedical informatics, 2018, vol. 83, pp. 103–111, DOI: 10.1016/j.jbi.2018.05.007.
28. Dong Y., Yang X. A hierarchical depression detection model based on vocal and emotional cues. Neurocomputing, 2021, vol. 441, pp. 279–290, DOI: 10.1016/j.neucom.2021.02.019.
29. Zheng W., Yan L., Wang F.-Y. Two birds with one stone: knowledge-embedded temporal convolutional transformer for depression detection and emotion recognition. IEEE Transactions on affective computing, 2023, vol. 14, no. 4, pp. 2595–2613, DOI: 10.1109/TAFFC.2023.3282704.
30. Degottex G., Kane J., Drugman T., et al. COVAREP – A collaborative voice analysis repository for speech technologies. 2014 IEEE International conference on acoustics, speech and signal processing (ICASSP), 2014, pp. 960–964, DOI: 10.1109/ICASSP.2014.6853739.
31. Kim J.C., Clements M.A. Formant-based feature extraction for emotion classification from speech. IEEE 2015 38th International conference on telecommunications and signal processing (TSP), 2015, pp. 477–481, DOI: 10.1109/TSP.2015.7296308.

32. Ringeval F., et al. AVEC'19. Proceedings of the 27th ACM international conference on multimedia. New York, NY, USA: ACM, 2019, pp. 2718–2719, DOI: 10.1145/3343031.3350550.
33. Tølbøll K.B. Linguistic features in depression: a meta-analysis. Journal of Language Works – Sprogvidenskabeligt Studentertidsskrift, 2019, vol. 4, no. 2, pp. 39–59, available at: https://tidsskrift.dk/lwo/article/view/117798
34. Al-Mosaiwi M., Johnstone T. In an absolute state: elevated use of absolutist words is a marker specific to anxiety, depression, and suicidal ideation. Clinical psychological science, 2018, vol. 6, no. 4, pp. 529–542, DOI: 10.1177/2167702617747074.
35. Bird S., Klein E., Loper E. Natural language processing with Python: analyzing text with the natural language toolkit. O'Reilly Media, Inc., 2009.
36. Malhotra A., Jindal R. XAI transformer-based approach for interpreting depressed and suicidal user behavior on online social networks. Cognitive systems research, 2023, p. 101186, DOI: 10.1016/j.cogsys.2023.101186.
37. Mitra V., et al. The SRI AVEC-2014 Evaluation System. Proceedings of the 4th international workshop on audio/visual emotion challenge, New York, NY, USA: ACM, 2014, pp. 93–101, DOI: 10.1145/2661806.2661818.
38. Shi X., et al. Convolutional LSTM network: a machine learning approach for precipitation nowcasting. 2015.
39. Amanat A., et al. Deep learning for depression detection from textual data. Electronics, 2022, vol. 11, no. 5, p. 676, DOI: 10.3390/electronics11050676.
40. Akter M.S., Shahriar H., Cuzzocrea A. A Trustable LSTM-autoencoder network for cyberbullying detection on social media using synthetic data, 2023, DOI: 10.48550/arXiv.2308.09722.
41. Lu J., et al. Prediction of depression severity based on transformer encoder and CNN Model. IEEE, 2022 13th International symposium on Chinese spoken language processing (ISCSLP), 2022, pp. 339–343, DOI: 10.1109/ISCSLP57327.2022.10038064.
42. Vaswani A., Shazeer N., Parmar N., et al. Attention is all you need, 2017, DOI:10.48550/arXiv.1706.03762 .
43. Kitaev N., Kaiser Ł., Levskaya A. Reformer: the efficient transformer, 2020, DOI: 10.48550/arXiv.2001.04451.
44. Mallol-Ragolta A., Zhao Z., Stappen L., et al. A Hierarchical attention network-based approach for depression detection from transcribed clinical interviews. Interspeech 2019, ISCA, 2019, pp. 221–225, DOI: 10.21437/Interspeech.2019-2036.
45. Zhang Y., He Y., Rong L., Ding Y. A hybrid model for depression detection with transformer and bi-directional long short-term memory. 2022 IEEE International conference on bioinformatics and biomedicine (BIBM), 2022, pp. 2727–2734, DOI: 10.1109/BIBM55620.2022.9995184.
46. Cheligeer C., et al. BERT-Based neural network for inpatient fall detection from electronic medical records: retrospective cohort study. JMIR medical informatics, 2024, vol. 12, p. e48995, DOI: 10.2196/48995.
47. Daru D., Surani H., Koladia H., et al. Depression detection using hybrid transformer networks, 2023, pp. 593–604, DOI: 10.1007/978-981-99-1414-2_44.
48. Tang S., Li C., Zhang P., Tang R. SwinLSTM: improving spatiotemporal prediction accuracy using swin transformer and LSTM, 2023, DOI: 10.48550/arXiv.2308.09891 .
49. Masci J., Meier U., Cireşan D., et al. Stacked convolutional auto-encoders for hierarchical feature extraction, 2011, pp. 52–59, DOI: 10.1007/978-3-642-21735-7_7.
50. Liang H., Sun X., Sun Y., et al. Text feature extraction based on deep learning: a review. EURASIP Journal on wireless communications and networking, 2017, vol. 2017, no. 1, p. 211, DOI: 10.1186/s13638-017-0993-1.
51. Chen Y., Zaki M.J. KATE. Proceedings of the 23rd ACM sigkdd international conference on knowledge discovery and data mining. New York, NY, USA: ACM, 2017, pp. 85–94, DOI: 10.1145/3097983.3098017.
52. İrsoy O., Alpaydın E. Unsupervised feature extraction with autoencoder trees. Neurocomputing, 2017, vol. 258, pp. 63–73, DOI: 10.1016/j.neucom.2017.02.075.
53. Mehrotra A., Musolesi M. Using autoencoders to automatically extract mobility features for predicting depressive states. Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies, 2018, vol. 2, no. 3, pp. 1–20, DOI: 10.1145/3264937.
54. Soares R.G.F. Effort estimation via text classification and autoencoders. IEEE 2018 International Joint Conference on Neural Networks (IJCNN), 2018, pp. 1–8, DOI: 10.1109/IJCNN.2018.8489030.
55. Gui T., Zhang Q., Zhu L., et al. Depression detection on social media with reinforcement learning, 2019, pp. 613–624, DOI: 10.1007/978-3-030-32381-3_49.
56. Che L., Yang X., Wang L. Text feature extraction based on stacked variational autoencoder. Microprocessors and microsystems, 2020, vol. 76, p. 103063, DOI: 10.1016/j.micpro.2020.103063.
57. Montero I., Pappas N., Smith N.A. Sentence bottleneck autoencoders from transformer language models. 2021, DOI: 10.48550/arXiv.2109.00055.
58. Rama K., Kumar P., Bhasker B. Deep autoencoders for feature learning with embeddings for recommendations: a novel recommender system solution. Neural computing and applications, 2021, vol. 33, no. 21, pp. 14167–14177, DOI: 10.1007/s00521-021-06065-9.

59. Ringeval F., et al. AVEC 2019 Workshop and challenge: state-of-mind, detecting depression with AI, and Cross-cultural affect recognition. Proceedings of the 9th International on audio/visual emotion challenge and workshop, New York, NY, USA: ACM, 2019, pp. 3–12, DOI: 10.1145/3347320.3357688.
60. Li M., Sun X., Wang M. Detecting depression with heterogeneous graph neural network in clinical interview transcript. IEEE Transactions on computational social systems, 2023, pp. 1–10, DOI: 10.1109/TCSS.2023.3263056.
61. Abbas M.A., et al. Novel Transformer based contextualized embedding and probabilistic features for depression detection from social media. IEEE Access, 2024, vol. 12, pp. 54087–54100, DOI: 10.1109/ACCESS.2024.3387695.
62. Hsu W.N., et al. HuBERT: self-supervised speech representation learning by masked prediction of hidden units. IEEE/ACM Transactions on audio, speech, and language processing, 2021, vol. 29, pp. 3451–3460, DOI: 10.1109/TASLP.2021.3122291.
63. Lam G., Dongyan H., Lin W. Context-aware deep learning for multi-modal depression detection. ICASSP 2019 - 2019 IEEE International conference on acoustics, speech and signal processing (ICASSP), 2019, pp. 3946–3950, DOI: 10.1109/ICASSP.2019.8683027.
64. Firoz N., Beresteneva O.G., Vladimirovich A.S., et al. Automated text-based depression detection using hybrid ConvLSTM and Bi-LSTM Model. 2023 Third International Conference on Artificial Intelligence and Smart Energy (ICAIS). IEEE, 2023, pp. 734–740, DOI: 10.1109/ICAIS56108.2023.10073683.
65. Verma B., Gupta S., Goel L. A Neural network based hybrid model for depression detection in Twitter. 2020, pp. 164–175, DOI: 10.1007/978-981-15-6634-9_16.
66. Vandana, Marriwala N., Chaudhary D. A hybrid model for depression detection using deep learning. Measurement: Sensors, 2023, vol. 25, p. 100587, DOI: 10.1016/j.measen.2022.100587.
67. Chen J., et al. IIFDD: Intra and inter-modal fusion for depression detection with multi-modal information from Internet of Medical Things. Information fusion, 2024, vol. 102, p. 102017, DOI: 10.1016/j.inffus.2023.102017.
68. Williamson J.R., et al. Detecting depression using vocal, facial and semantic communication cues. Proceedings of the 6th International Workshop on Audio/visual emotion challenge, New York, NY, USA: ACM, 2016, pp. 11–18, DOI: 10.1145/2988257.2988263.
69. Al Hanai T., Ghassemi M., Glass J. Detecting depression with audio/text sequence modeling of interviews. interspeech 2018. ISCA, 2018, pp. 1716–1720, DOI: 10.21437/Interspeech.2018-2522.
70. Rodrigues Makiuchi M., Warnita T., et al. Multimodal fusion of BERT-CNN and gated CNN representations for depression detection. Proceedings of the 9th International on audio/visual emotion challenge and workshop. New York, NY, USA: ACM, 2019, pp. 55–63, DOI: 10.1145/3347320.3357694.
71. Vandana, Marriwala N., Chaudhary D. A hybrid model for depression detection using deep learning. Measurement: Sensors, 2023, vol. 25, p. 100587, DOI: 10.1016/j.measen.2022.100587.
72. Das A.K., Naskar R. A deep learning model for depression detection based on MFCC and CNN generated spectrogram features. Biomedical signal processing and control, 2024, vol. 90, p. 105898, DOI: 10.1016/j.bspc.2023.105898.
73. Bertl M., et al. Evaluation of deep learning-based depression detection using medical claims data. Artificial Intelligence in medicine, 2024, vol. 147, p. 102745, DOI: 10.1016/j.artmed.2023.102745.
74. Mo H., et al. A Multimodal data-driven framework for anxiety screening. IEEE Transactions on instrumentation and measurement, 2024, vol. 73, pp. 1–13, DOI: 10.1109/TIM.2024.3352713.
75. Iyortsuun N.K., et al. Additive cross-modal attention network (ACMA) for depression detection based on audio and textual features. IEEE Access, 2024, vol. 12, pp. 20479–20489, DOI: 10.1109/ACCESS.2024.3362233.
76. Fan C., et al. Light-weight residual convolution-based capsule network for EEG emotion recognition. Advanced Engineering Informatics, 2024, vol. 61, p. 102522, DOI: 10.1016/j.aei.2024.102522.
77. Fan H., et al. Transformer-based multimodal feature enhancement networks for multimodal depression detection integrating video, audio and remote photoplethysmograph signals. Information Fusion, 2024, vol. 104, p. 102161, DOI: 10.1016/j.inffus.2023.102161.
78. Artstein R., et al. SimSensei Kiosk: A virtual human interviewer for healthcare decision support, 2014.
79. Manning C., Schutze H. Foundations of Statistical Natural Language Processing. MIT Press, 1999.
80. Hotho A., Nürnberger A., Paaß G. A Brief Survey of Text Mining. Journal for Language Technology and Computational Linguistics, 2005, vol. 20, no. 1, pp. 19–62, DOI: 10.21248/jlcl.20.2005.68.
81. Lin L., Chen X., Shen Y., Zhang L. Towards automatic depression detection: A BiLSTM/1D CNN-Based Model. Applied sciences, 2020, vol. 10, no. 23, p. 8701, DOI: 10.3390/app10238701.
82. Sun G., Zhao S., Zou B., An Y. Multimodal depression detection using a deep feature fusion network. Third International Conference on Computer Science and Communication Technology (ICCSCT 2022). SPIE, 2022, p. 269, DOI: 10.1117/12.2662620.

83. Bennetot A., Laurent J.-L., Chatila R., et al. Towards Explainable Neural-Symbolic Visual Reasoningm 2019.
84. Ye J., Zhang J., Shan H. DepMamba: Progressive fusion mamba for multimodal depression detection. ICASSP 2025 – 2025 IEEE International conference on acoustics, speech and signal processing (ICASSP), Hyderabad, India, 2025, pp. 1–5, DOI: 10.1109/ICASSP49660.2025.10889975.

Неда Фироз. *Научный сотрудник Томского государственного университета. Научные интересы лежат в области искусственного интеллекта и машинного обучения, с особым акцентом на оптимизацию диагностики депрессии с использованием передовых методов глубокого обучения. AuthorID: 1237549, SPIN-код: 4497-4847, ORCID 0000-0003-4696-2072, nedafiroz1910@gamil.com.*

Берестнева Ольга Григорьевна. *Доктор технических наук, профессор кафедры информационных технологий Томского политехнического университета. AuthorID: 112998, SPIN: 8026-4116, ORCID 0000-0002-4243-0637, ogb6@yandex.ru, 634050, Россия, Томск-50, проспект Ленина, 30.*

Аксенов Сергей Владимирович. *Кандидат технических наук, доцент кафедры информационных технологий Томского политехнического университета (ТПУ). Область научных интересов: разработка интеллектуальных систем для анализа многомерных данных, технологий распараллеливания вычислений и компьютерного зрения. AuthorID: 505275, SPIN-код: 2229-4552. axyonov@tpu.ru.*

UDC 004.81
DOI:10.25729/ESI.2025.40.4.005

Multimodal depression detection using Multistream Mood Insight Encoder (MMIE)

Neda Firoz[1], Olga G. Berestneva[2], Sergey V. Aksenov[2]

[1]Tomsk State University,
Russia, Tomsk, *nedafiroz1910@gmail.com*
[2]Tomsk Polytechnic University,
Russia, Tomsk

Abstract. The global surge in the prevalence of depression, characterized by persistent sadness, disinterest, and decreased functioning, highlights the shortcomings of prevailing diagnostic and treatment paradigms. This underscores the urgent need for enhanced interventions, given the inherent limitations of traditional approaches to diagnosing depression. Recent advances in artificial intelligence applications have sparked growing interest in the development of automated depression diagnostic systems among emotion computing experts. The emergence of large-scale language models, such as BERT and its derivatives, for text-based depression detection demonstrates the need for multimodal approaches that integrate text and audio data to achieve more accurate diagnosis. Here, we explore the capabilities of existing large-scale language models and present a proposed multi-stream model, the Multi-Stream Mood Insight Encoder (MMIE). MMIE is designed to seamlessly utilize integrated text and audio data streams with processing capabilities via the Reformer encoder. As part of this concept, linguistic features such as absolutist words and first-person pronouns were incorporated into the Reformer encoder. This holistic approach facilitated a comprehensive analysis of a person's mood and emotional state. Experiments demonstrated that the ClinicalBERT language model outperformed the proposed binary depression classification model. Subsequently, the sigmoid values of the Reformer model were used to diagnose depression. Using the proposed model, experiments were conducted on the DAIC-WOZ dataset. The results showed significant improvements, demonstrating an F1 of 0.9538 for classification, an MAE of 3.42, and an RMSE of 4.64 for regression compared to state-of-the-art methods. These results demonstrate the effectiveness of the proposed model in facilitating the diagnosis of depression.

Keywords: audio, clinical analysis, depression detection, LLMs, Reformer, MMIE

Acknowledgements: The authors express their gratitude to the DAIC_WOZ dataset license, which significantly facilitated their research efforts. They are also grateful to the Tomsk State University Library for providing the authors with access to a wide range of valuable resources. This research was supported by a grant from the Government of the Russian Federation (Agreement No. 075-15-2025-009 dated February 28, 2025).

References

1. Institute of Health Metrics and Evaluation. Global health data exchange (GHDx), 2021.
2. Twenge J.M., Cooper A.B., Joiner T.E., et al. Age, period, and cohort trends in mood disorder indicators and suicide-related outcomes in a nationally representative dataset, 2005–2017. Journal of abnormal psychology, 2019, vol. 128, no. 3, pp. 185–199, DOI: 10.1037/abn0000410.
3. Diagnostic and statistical manual of mental disorders. 5th ed. American psychiatric association, 2013.
4. Nestler E.J., Barrot M., DiLeone R.J., et al. Neurobiology of Depression. Neuron, 2002, vol. 34, no. 1, pp. 13–25, DOI: 10.1016/S0896-6273(02)00653-0.
5. Cohen K. Absolutist thinking and depression, 2019.
6. Blazer D.G. Psychiatry and the oldest old. American journal of psychiatry, 2000, vol. 157, no. 12, pp. 1915–1924, DOI: 10.1176/appi.ajp.157.12.1915.
7. Zarate C.A., et al. A Randomized Trial of an N-methyl-D-aspartate Antagonist in treatment-resistant major depression. Archives of general psychiatry, 2006, vol. 63, no. 8, p. 856, DOI: 10.1001/archpsyc.63.8.856.
8. Chattopadhyay S. A neuro-fuzzy approach for the diagnosis of depression. Applied Computing and Informatics, 2017, vol. 13, no. 1, pp. 10–18, DOI: 10.1016/j.aci.2014.01.001.
9. Joshi M.L., Kanoongo N. Depression detection using emotional artificial intelligence and machine learning: A closer review. Materials Today: Proceedings, 2022, vol. 58, pp. 217–226, DOI: 10.1016/j.matpr.2022.01.467.
10. Miao X., Li Y., Wen M., Liu Y., et al. Fusing features of speech for depression classification based on higher-order spectral analysis. Speech Communication, 2022, vol. 143, pp. 46–56, DOI: 10.1016/j.specom.2022.07.006.
11. Ilias L., Mouzakitis S., Askounis D. Calibration of transformer-based models for identifying stress and depression in social media. IEEE Transactions on computational social systems, 2023, pp. 1–12, DOI: 10.1109/TCSS.2023.3283009.
12. Oureshi S.A., Dias G., Saha S., Hasanuzzaman M. Gender-aware estimation of depression severity level in a multimodal setting. 2021 International joint conference on neural networks (IJCNN), IEEE, 2021, pp. 1–8, DOI: 10.1109/IJCNN52387.2021.9534330.
13. Firoz N., Beresteneva O.G., Vladimirovich A.S., et al. Automated text-based depression detection using hybrid ConvLSTM and Bi-LSTM model. 2023 Third International conference on artificial intelligence and smart energy (ICAIS). IEEE, 2023, pp. 734–740, DOI: 10.1109/ICAIS56108.2023.10073683.
14. Ahmad Wani M., ELaffendi M.A., Shakil K.A., et al. Depression screening in humans with AI and deep learning techniques. IEEE Transactions on computational social systems, 2023, vol. 10, no. 4, pp. 2074–2089, DOI: 10.1109/TCSS.2022.3200213.
15. Mazumdar H., Chakraborty C., Sathvik M., et al. GPTFX: A Novel GPT-3 based framework for mental health detection and explanations. IEEE Journal of biomedical and health informatics, 2023, pp. 1–8, DOI: 10.1109/JBHI.2023.3328350.
16. Meng Q., Catchpoole D., Skillicom D., et al. Relational autoencoder for feature extraction. 2017 International Joint conference on neural networks (IJCNN), IEEE, 2017, pp. 364–371, DOI: 10.1109/IJCNN.2017.7965877.
17. Devlin J., Chang M.-W., Lee K., Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding, 2018, DOI: 10.48550/arXiv.1810.04805.
18. Ji S., Zhang T., Ansari L., Fu J., et al. MentalBERT: publicly available pretrained language models for mental healthcare. Proceedings of the language resources and evaluation conference (LREC), 2022, DOI: 10.48550/arXiv.2110.15621.
19. Vajre V., Naylor M., Kamath U., et al. PsychBERT: a mental health language model for social media mental health behavioral analysis. 2021 IEEE International conference on bioinformatics and biomedicine (BIBM), 2021, pp. 1077–1082, DOI: 10.1109/BIBM52615.2021.9669469.
20. Huang K., Altosaar J., Ranganath R. ClinicalBERT: modeling clinical notes and predicting hospital readmission, 2019.
21. Saha T., Ramesh Jayashree S., Saha S., et al. BERT-caps: a transformer-based capsule network for tweet act classification. IEEE Transactions on computational social systems, 2020, vol. 7, no. 5, pp. 1168–1179, DOI: 10.1109/TCSS.2020.3014128.
22. Saha T., Reddy S.M., Saha S., et al. Mental health disorder identification from motivational conversations. IEEE Transactions on computational social systems, 2023, vol. 10, no. 3, pp. 1130–1139, DOI: 10.1109/TCSS.2022.3143763.
23. Li M. , et al. TrOCR: Transformer-based optical character recognition with pre-trained models. Proceedings of the AAAI conference on Artificial Intelligence, 2023, vol. 37, no. 11, pp. 13094–13102, DOI: 10.1609/aaai.v37i11.26538.
24. Anindyaputri N.A., Girsang A.S. A Comparative study of deep learning models for detecting depressive disorder in tweets. Journal of system and management sciences, 2024, vol. 14, no. 3, DOI: 10.33168/JSMS.2024.0318.

25. Flint A.J., Black S.E., Campbell-Taylor I., et al. Abnormal speech articulation, psychomotor retardation, and subcortical dysfunction in major depression. Journal of psychiatric research, 1993, vol. 27, no. 3, pp. 309–319, DOI: 10.1016/0022-3956(93)90041-Y.
26. Korszun A. Facial pain, depression and stress – connections and directions. Journal of oral pathology & medicine, 2002, vol. 31, no. 10, pp. 615–619, DOI: 10.1034/j.1600-0714.2002.00091.x.
27. He L., Cao C. Automated depression analysis using convolutional neural networks from speech. Journal of biomedical informatics, 2018, vol. 83, pp. 103–111, DOI: 10.1016/j.jbi.2018.05.007.
28. Dong Y., Yang X. A hierarchical depression detection model based on vocal and emotional cues. Neurocomputing, 2021, vol. 441, pp. 279–290, DOI: 10.1016/j.neucom.2021.02.019.
29. Zheng W., Yan L., Wang F.-Y. Two birds with one stone: knowledge-embedded temporal convolutional transformer for depression detection and emotion recognition. IEEE Transactions on affective computing, 2023, vol. 14, no. 4, pp. 2595–2613, DOI: 10.1109/TAFFC.2023.3282704.
30. Degottex G., Kane J., Drugman T., et al. COVAREP – A collaborative voice analysis repository for speech technologies. 2014 IEEE International conference on acoustics, speech and signal processing (ICASSP), 2014, pp. 960–964, DOI: 10.1109/ICASSP.2014.6853739.
31. Kim J.C., Clements M.A. Formant-based feature extraction for emotion classification from speech. IEEE 2015 38th International conference on telecommunications and signal processing (TSP), 2015, pp. 477–481, DOI: 10.1109/TSP.2015.7296308.
32. Ringeval F., et al. AVEC'19. Proceedings of the 27th ACM international conference on multimedia. New York, NY, USA: ACM, 2019, pp. 2718–2719, DOI: 10.1145/3343031.3350550.
33. Tølbøll K.B. Linguistic features in depression: a meta-analysis. Journal of Language Works – Sprogvidenskabeligt Studentertidsskrift, 2019, vol. 4, no. 2, pp. 39–59, available at: https://tidsskrift.dk/lwo/article/view/117798
34. Al-Mosaiwi M., Johnstone T. In an absolute state: elevated use of absolutist words is a marker specific to anxiety, depression, and suicidal ideation. Clinical psychological science, 2018, vol. 6, no. 4, pp. 529–542, DOI: 10.1177/2167702617747074.
35. Bird S., Klein E., Loper E. Natural language processing with Python: analyzing text with the natural language toolkit. O'Reilly Media, Inc., 2009.
36. Malhotra A., Jindal R. XAI transformer-based approach for interpreting depressed and suicidal user behavior on online social networks. Cognitive systems research, 2023, p. 101186, DOI: 10.1016/j.cogsys.2023.101186.
37. Mitra V., et al. The SRI AVEC-2014 Evaluation System. Proceedings of the 4th international workshop on audio/visual emotion challenge, New York, NY, USA: ACM, 2014, pp. 93–101, DOI: 10.1145/2661806.2661818.
38. Shi X., et al. Convolutional LSTM network: a machine learning approach for precipitation nowcasting. 2015.
39. Amanat A., et al. Deep learning for depression detection from textual data. Electronics, 2022, vol. 11, no. 5, p. 676, DOI: 10.3390/electronics11050676.
40. Akter M.S., Shahriar H., Cuzzocrea A. A Trustable LSTM-autoencoder network for cyberbullying detection on social media using synthetic data, 2023, DOI: 10.48550/arXiv.2308.09722.
41. Lu J., et al. Prediction of depression severity based on transformer encoder and CNN Model. IEEE, 2022 13th International symposium on Chinese spoken language processing (ISCSLP), 2022, pp. 339–343, DOI: 10.1109/ISCSLP57327.2022.10038064.
42. Vaswani A., Shazeer N., Parmar N., et al. Attention is all you need, 2017, DOI:10.48550/arXiv.1706.03762 .
43. Kitaev N., Kaiser Ł., Levskaya A. Reformer: the efficient transformer, 2020, DOI: 10.48550/arXiv.2001.04451.
44. Mallol-Ragolta A., Zhao Z., Stappen L., et al. A Hierarchical attention network-based approach for depression detection from transcribed clinical interviews. Interspeech 2019, ISCA, 2019, pp. 221–225, DOI: 10.21437/Interspeech.2019-2036.
45. Zhang Y., He Y., Rong L., Ding Y. A hybrid model for depression detection with transformer and bi-directional long short-term memory. 2022 IEEE International conference on bioinformatics and biomedicine (BIBM), 2022, pp. 2727–2734, DOI: 10.1109/BIBM55620.2022.9995184.
46. Cheligeer C., et al. BERT-Based neural network for inpatient fall detection from electronic medical records: retrospective cohort study. JMIR medical informatics, 2024, vol. 12, p. e48995, DOI: 10.2196/48995.
47. Daru D., Surani H., Koladia H., et al. Depression detection using hybrid transformer networks, 2023, pp. 593–604, DOI: 10.1007/978-981-99-1414-2_44.
48. Tang S., Li C., Zhang P., Tang R. SwinLSTM: improving spatiotemporal prediction accuracy using swin transformer and LSTM, 2023, DOI: 10.48550/arXiv.2308.09891 .
49. Masci J., Meier U., Cireşan D., et al. Stacked convolutional auto-encoders for hierarchical feature extraction, 2011, pp. 52–59, DOI: 10.1007/978-3-642-21735-7_7.
50. Liang H., Sun X., Sun Y., et al. Text feature extraction based on deep learning: a review. EURASIP Journal on wireless communications and networking, 2017, vol. 2017, no. 1, p. 211, DOI: 10.1186/s13638-017-0993-1.

51. Chen Y., Zaki M.J. KATE. Proceedings of the 23rd ACM sigkdd international conference on knowledge discovery and data mining. New York, NY, USA: ACM, 2017, pp. 85–94, DOI: 10.1145/3097983.3098017.
52. İrsoy O., Alpaydın E. Unsupervised feature extraction with autoencoder trees. Neurocomputing, 2017, vol. 258, pp. 63–73, DOI: 10.1016/j.neucom.2017.02.075.
53. Mehrotra A., Musolesi M. Using autoencoders to automatically extract mobility features for predicting depressive states. Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies, 2018, vol. 2, no. 3, pp. 1–20, DOI: 10.1145/3264937.
54. Soares R.G.F. Effort estimation via text classification and autoencoders. IEEE 2018 International Joint Conference on Neural Networks (IJCNN), 2018, pp. 1–8, DOI: 10.1109/IJCNN.2018.8489030.
55. Gui T., Zhang Q., Zhu L., et al. Depression detection on social media with reinforcement learning, 2019, pp. 613–624, DOI: 10.1007/978-3-030-32381-3_49.
56. Che L., Yang X., Wang L. Text feature extraction based on stacked variational autoencoder. Microprocessors and microsystems, 2020, vol. 76, p. 103063, DOI: 10.1016/j.micpro.2020.103063.
57. Montero I., Pappas N., Smith N.A. Sentence bottleneck autoencoders from transformer language models. 2021, DOI: 10.48550/arXiv.2109.00055.
58. Rama K., Kumar P., Bhasker B. Deep autoencoders for feature learning with embeddings for recommendations: a novel recommender system solution. Neural computing and applications, 2021, vol. 33, no. 21, pp. 14167–14177, DOI: 10.1007/s00521-021-06065-9.
59. Ringeval F., et al. AVEC 2019 Workshop and challenge: state-of-mind, detecting depression with AI, and Cross-cultural affect recognition. Proceedings of the 9th International on audio/visual emotion challenge and workshop, New York, NY, USA: ACM, 2019, pp. 3–12, DOI: 10.1145/3347320.3357688.
60. Li M., Sun X., Wang M. Detecting depression with heterogeneous graph neural network in clinical interview transcript. IEEE Transactions on computational social systems, 2023, pp. 1–10, DOI: 10.1109/TCSS.2023.3263056.
61. Abbas M.A., et al. Novel Transformer based contextualized embedding and probabilistic features for depression detection from social media. IEEE Access, 2024, vol. 12, pp. 54087–54100, DOI: 10.1109/ACCESS.2024.3387695.
62. Hsu W.N., et al. HuBERT: self-supervised speech representation learning by masked prediction of hidden units. IEEE/ACM Transactions on audio, speech, and language processing, 2021, vol. 29, pp. 3451–3460, DOI: 10.1109/TASLP.2021.3122291.
63. Lam G., Dongyan H., Lin W. Context-aware deep learning for multi-modal depression detection. ICASSP 2019 - 2019 IEEE International conference on acoustics, speech and signal processing (ICASSP), 2019, pp. 3946–3950, DOI: 10.1109/ICASSP.2019.8683027.
64. Firoz N., Beresteneva O.G., Vladimirovich A.S., et al. Automated text-based depression detection using hybrid ConvLSTM and Bi-LSTM Model. 2023 Third International Conference on Artificial Intelligence and Smart Energy (ICAIS). IEEE, 2023, pp. 734–740, DOI: 10.1109/ICAIS56108.2023.10073683.
65. Verma B., Gupta S., Goel L. A Neural network based hybrid model for depression detection in Twitter. 2020, pp. 164–175, DOI: 10.1007/978-981-15-6634-9_16.
66. Vandana, Marriwala N., Chaudhary D. A hybrid model for depression detection using deep learning. Measurement: Sensors, 2023, vol. 25, p. 100587, DOI: 10.1016/j.measen.2022.100587.
67. Chen J., et al. IIFDD: Intra and inter-modal fusion for depression detection with multi-modal information from Internet of Medical Things. Information fusion, 2024, vol. 102, p. 102017, DOI: 10.1016/j.inffus.2023.102017.
68. Williamson J.R., et al. Detecting depression using vocal, facial and semantic communication cues. Proceedings of the 6th International Workshop on Audio/visual emotion challenge, New York, NY, USA: ACM, 2016, pp. 11–18, DOI: 10.1145/2988257.2988263.
69. Al Hanai T., Ghassemi M., Glass J. Detecting depression with audio/text sequence modeling of interviews. interspeech 2018. ISCA, 2018, pp. 1716–1720, DOI: 10.21437/Interspeech.2018-2522.
70. Rodrigues Makiuchi M., Warnita T., et al. Multimodal fusion of BERT-CNN and gated CNN representations for depression detection. Proceedings of the 9th International on audio/visual emotion challenge and workshop. New York, NY, USA: ACM, 2019, pp. 55–63, DOI: 10.1145/3347320.3357694.
71. Vandana, Marriwala N., Chaudhary D. A hybrid model for depression detection using deep learning. Measurement: Sensors, 2023, vol. 25, p. 100587, DOI: 10.1016/j.measen.2022.100587.
72. Das A.K., Naskar R. A deep learning model for depression detection based on MFCC and CNN generated spectrogram features. Biomedical signal processing and control, 2024, vol. 90, p. 105898, DOI: 10.1016/j.bspc.2023.105898.
73. Bertl M., et al. Evaluation of deep learning-based depression detection using medical claims data. Artificial Intelligence in medicine, 2024, vol. 147, p. 102745, DOI: 10.1016/j.artmed.2023.102745.

74. Mo H., et al. A Multimodal data-driven framework for anxiety screening. IEEE Transactions on instrumentation and measurement, 2024, vol. 73, pp. 1–13, DOI: 10.1109/TIM.2024.3352713.
75. Iyortsuun N.K., et al. Additive cross-modal attention network (ACMA) for depression detection based on audio and textual features. IEEE Access, 2024, vol. 12, pp. 20479–20489, DOI: 10.1109/ACCESS.2024.3362233.
76. Fan C., et al. Light-weight residual convolution-based capsule network for EEG emotion recognition. Advanced Engineering Informatics, 2024, vol. 61, p. 102522, DOI: 10.1016/j.aei.2024.102522.
77. Fan H., et al. Transformer-based multimodal feature enhancement networks for multimodal depression detection integrating video, audio and remote photoplethysmograph signals. Information Fusion, 2024, vol. 104, p. 102161, DOI: 10.1016/j.inffus.2023.102161.
78. Artstein R., et al. SimSensei Kiosk: A virtual human interviewer for healthcare decision support, 2014.
79. Manning C., Schutze H. Foundations of Statistical Natural Language Processing. MIT Press, 1999.
80. Hotho A., Nürnberger A., Paaß G. A Brief Survey of Text Mining. Journal for Language Technology and Computational Linguistics, 2005, vol. 20, no. 1, pp. 19–62, DOI: 10.21248/jlcl.20.2005.68.
81. Lin L., Chen X., Shen Y., Zhang L. Towards automatic depression detection: A BiLSTM/1D CNN-Based Model. Applied sciences, 2020, vol. 10, no. 23, p. 8701, DOI: 10.3390/app10238701.
82. Sun G., Zhao S., Zou B., An Y. Multimodal depression detection using a deep feature fusion network. Third International Conference on Computer Science and Communication Technology (ICCSCT 2022). SPIE, 2022, p. 269, DOI: 10.1117/12.2662620.
83. Bennetot A., Laurent J.-L., Chatila R., et al. Towards Explainable Neural-Symbolic Visual Reasoningm 2019.
84. Ye J., Zhang J., Shan H. DepMamba: Progressive fusion mamba for multimodal depression detection. ICASSP 2025 – 2025 IEEE International conference on acoustics, speech and signal processing (ICASSP), Hyderabad, India, 2025, pp. 1–5, DOI: 10.1109/ICASSP49660.2025.10889975.

Firoz Neda. *Research fellow, Tomsk State University. Her research interests lie in artificial intelligence and machine learning, with a particular focus on optimizing depression diagnostics using advanced deep learning methods. AuthorID: 112998, SPIN: 8026-4116, ORCID 0000-0003-4696-2072, nedafiroz1910@gamil.com.*

Berestneva Olga Grigoryevna. *Doctor of Engineering Sciences, Professor, Department of Information Technology, Tomsk Polytechnic University. AuthorID: 112998, SPIN: 8026-4116, ORCID 0000-0002-4243-0637, ogb6@yandex.ru, 634050, Russia, Tomsk-50, 30 Lenin Avenue.*

Aksenov Sergey Vladimirovich. *Candidate of Technical Sciences, Associate Professor in the Information Technology Department at Tomsk Polytechnic University (TPU). His research interests include developing intelligent systems for multidimensional data analysis, parallel computing technologies, and computer vision. AuthorID: 505275, SPIN-код: 2229-4552. axyonov@tpu.ru.*

Статья поступила в редакцию 01.07.2025; одобрена после рецензирования 01.10.2025; принята к публикации 09.11.2025.

The article was submitted 07/01/2025; approved after reviewing 10/01/2025; accepted for publication 11/09/2025.