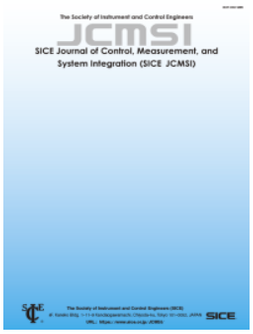


SICE Journal of Control, Measurement, and System Integration

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/tmsi20

# Enhancing early depression detection with AI: a comparative use of NLP models

Bakir Hadzic, Parvez Mohammed, Michael Danner, Julia Ohse, Yihong Zhang, Youssef Shiban & Matthias Rätsch

**To cite this article:** Bakir Hadzic, Parvez Mohammed, Michael Danner, Julia Ohse, Yihong Zhang, Youssef Shiban & Matthias Rätsch (2024) Enhancing early depression detection with AI: a comparative use of NLP models, SICE Journal of Control, Measurement, and System Integration, 17:1, 135-143, DOI: 10.1080/18824889.2024.2342624

**To link to this article:** https://doi.org/10.1080/18824889.2024.2342624



Published online: 23 Apr 2024.

SICE JOURNAL OF CONTROL, MEASUREMENT, AND SYSTEM INTEGRATION
2024, VOL. 17, NO. 1, 135–143
https://doi.org/10.1080/18824889.2024.2342624



# Enhancing early depression detection with AI: a comparative use of NLP models

Bakir Hadzic [a†], Parvez Mohammed [a†], Michael Danner [b], Julia Ohse [c], Yihong Zhang [d], Youssef Shiban [c] and Matthias Rätsch [a]

[a]ViSiR, Reutlingen University, Reutlingen, Germany; [b]CVSSP, University of Surrey, Guildford, UK; [c]Private University of Applied Sciences, Göttingen, Germany; [d]College of Information Science and Technology, Donghua University, Shanghai, People's Republic of China

**ABSTRACT**

One of the most underdiagnosed medical conditions worldwide is depression. It has been demonstrated that the current classical procedures for early detection of depression are insufficient, which emphasizes the importance of seeking a more efficient approach to overcome this challenge. One of the most promising opportunities is arising in the field of Artificial Intelligence as AI-based models could have the capacity to offer a fast, widely accessible, unbiased and efficient method to address this problem. In this paper, we compared three natural language processing models, namely, BERT, GPT-3.5 and GPT-4 on three different datasets. Our findings show that different levels of efficacy are shown by fine-tuned BERT, GPT-3.5, and GPT-4 in identifying depression from textual data. By comparing the models on the metrics such as accuracy, precision, and recall, our results have shown that GPT-4 outperforms both BERT and GPT-3.5 models, even without previous fine-tuning, showcasing its enormous potential to be utilized for automated depression detection on textual data. In the paper, we present newly introduced datasets, fine-tuning and model testing processes, while also addressing limitations and discussing further considerations for future research.

**ARTICLE HISTORY**
Received 31 October 2023
Revised 1 February 2024
Accepted 9 April 2024

**KEYWORDS**
Mental health; depression detection; LLM; BERT; GPT-4

## 1. Introduction

According to the World Health Organization's most recent report on mental health [1], one in eight people worldwide lives with some sort of mental health disorder. For decades, there has been a stagnation in research on how to find new and more effective ways to detect and diagnose depression [2]. Consequently, currently available techniques for detecting depression are insufficiently effective. Unfortunately, almost two-thirds of those who require it do not obtain mental health care, primarily as a result of limitations to reaching mental health professionals, stigmatization, high costs, or extensive waiting lists. All of these factors contribute to the vast majority of persons with mental health disorders remaining undetected [3]. This issue is particularly evident in rural areas where mental health systems are not up to standard and the general population lacks access to proper mental health treatments. For example, more than 123 million Americans reside in mental health professional shortage areas [4]. The ability of a therapist to get the necessary diagnostic information from a patient, who mostly has a diminished outlook and motivation, depends mainly on their competence and experience making the process of depression detection quite challenging and time-consuming processes [5].

Due to the reasons previously mentioned, there is an immense demand for novel strategies to address these issues and improve existing methods so that they may become automated, faster, non-invasive, less costly, as well as accessible to larger populations. Great possibilities for that are arising from the area of artificial intelligence (AI) and machine learning where various methods for that purpose are currently being tested and developed. Some of them are presented under the related work section of this paper. Our approach has the ultimate goal of developing an AI-based method for early depression detection by analysing textual data. But at the same time, it is important to notice that this model should not ever be used to make a final clinical diagnosis, but rather as a supporting tool for mental health experts to detect individuals during the early onset of depressive symptoms. Its swift use and easy accessibility should offer alternatives for many individuals who do not have access to proper opportunities to contact mental health experts. Feedback from this system should serve as a first indication for them that they have to seek professional help from trained experts.

This paper should be considered as a comprehensive work of our previous paper [6] complementing it with significant novelties such as the inclusion of the GPT-4 model that was not accessible for research purposes up

**CONTACT** Matthias Rätsch matthias.raetsch@reutlingen-university.de
[†]These authors contributed equally to this work.

to this date and novel test data set. This novel dataset was collected through the interview method to expose our models to more rigorous tests using real-world data, but at the same time to compare the validity of the simulated data that was proposed in the previous work [6]. Additionally, we used this dataset to compare the performance of all models presented in previous work [6].

The related work, methodology, results, discussion and conclusion parts are presented in the following chapters.

## 2. Related work

Depressive symptoms can express themselves through both verbal and nonverbal channels. These channels can encompass a wide range of modalities, such as voice, prosody, speech content, facial expressions, body postures, and other behavioural indicators [7,8]. Such modalities can serve as sources of information that reflect the emotional state of an individual. By leveraging the detection of depression across multiple modalities, more robust and accurate insights can be extracted from a multi-dimensional perspective [9]. This can be achieved through various methods, including the use of audio recordings, where the voice, prosody and spoken text content can be utilized as data sources. Alternatively, depression can be detected through the recording of facial expressions, head postures, or gaze directions [7]. This chapter outlines the state-of-the-art developments in the area of text-based and audio-based depression detection, leveraging deep learning methodologies. Furthermore, it highlights the relevance of the advancements in the field of AI-based dialogue systems, especially in the area of large language models.

### 2.1. Text-based

One approach for detecting depression is to analyse speech and language patterns in individuals. Specifically, the text-based recognition of depression involves extracting and transcribing audio data to reveal the content of natural language [7]. This content can then be used to extract many pieces of information related to the emotional state of depressed individuals [9]. For the classification of this data, feature vectors are extracted and learned by a deep learning model [10]. In recent years, several algorithms, including convolutional neural networks (CNNs), long short-term memory (LSTM) networks, and transformers, have been utilized for text-based depression recognition [10,11]. Among these, transformer-based deep learning models have shown exceptional performance in the context recognition of natural language [12]. Consequently, such models have become a promising research direction in the field of depression diagnosis and treatment [10,11,13–15]. Toto et al. [13] present an example of a smartphone application named EMU3 that aims to detect depression using a multimodal speech classification system called AudiBERT. The proposed AudiBERT model utilizes dual self-attention mechanisms to establish the relationship between the structures of text and audio recordings. The model is trained and tested on the DAIC-WOZ dataset using BERT-based text recognition. The results demonstrate the effectiveness of AudiBERT in accurately detecting depression using both audio and text features. The proposed framework has promising potential for developing scalable, accessible, and cost-effective tools for the early detection and treatment of depression.

### 2.2. Audio

Deep learning models are capable of automatically extracting and learning complex patterns and relationships from audio data, which enables more accurate and reliable detection of depression [16]. The audio features used for training these models include pitch, intensity, and spectral features, which can be extracted from audio recordings using techniques such as Mel-frequency cepstral coefficients (MFCCs) [17], pitch analysis, and spectral feature analysis. The majority of current speech processing techniques first divide speech into brief (10–20 ms) frames before extracting low-level descriptors (such as spectral, prosodic, and glottal features) and high-level representations of those features (such as statistical functionals, such as mean and percentiles), vocal tract coordination (VTC) features, i-vectors, and Fisher vectors) [18]. Transformer-based methods have shown promising results in audio feature extraction and have become increasingly popular in recent years. McGinnis et al. [19] present an approach that is capable of recognising if children have internalized disorders like depression or anxiety. For that they need only 3 minutes of recorded speech and the accuracy of their approach is around 80%, which outperforms clinical thresholds on parent-reported child symptoms. Despite the fact that approaches using audio-based inputs have achieved quite promising results, Baileys and Plumbley [20] point out in their paper the presence of gender bias in depression detection models that utilize audio features. The study found that existing depression detection models trained on audio data exhibit significant gender bias, with higher accuracy for detecting depression in female voices compared to male voices. Additionally, the authors in their paper suggest potential solutions to mitigate the bias with a few different approaches. One proposed solution involves balancing the gender ratio in the training data used for the model. Another proposed solution is to use more advanced models that are capable of capturing subtle differences in acoustic features between male and female voices. Overall,

the paper highlights the importance of addressing gender bias in depression detection models and emphasizes the need for more diverse and representative training data to improve the accuracy and fairness of these models.

### 2.3. Large-scale language models

The development of large-scale language models (LLMs) has revolutionized the field of natural language processing (NLP). This has led to the emergence of several chatbots, based on LMMs, such as ChatGPT (based on LLM GPT 3.5/4.0 proposed by OpenAI/Microsoft [21]), BARD (based on LLM LaMDa, proposed by Google [22]), ERNIE (based on LLM ERNIE 2.3/3.0 [23] proposed by Baidu), or Dalai, Alpaca and others (based on LLM LLaMA proposed by Meta, Facebook [24]). These LLMs, based on the transformer architecture, have demonstrated significant advances in conversational AI and information retrieval.

It excels at generating coherent and contextually relevant responses, even for ambiguous queries. Chatbots based on LLMs are used in many fields of research and even in several everyday tasks. ChatGPT, an iteration of OpenAI's GPT series, has achieved outstanding performance in a wide range of tasks, such as translation, summarising, and question-answering. Its ability to generalize from few-shot learning examples is a testament to the model's adaptability. In parallel, Baidu's ERNIE model, which employs continual pre-training and knowledge distillation techniques, has shown impressive results in Chinese NLP tasks and achieved state-of-the-art performance on various benchmarks. Lastly, Meta's LLaMA model, utilising a combination of unsupervised and supervised learning, has been designed to handle low-resource languages and multi-modal data effectively. Each of these models offers unique contributions to the advancement of language understanding and generation, paving the way for more sophisticated AI applications. In the coming years, this recent progress in the LLM systems has the potential to revolutionize the field of AI research and to bring AI into the everyday life of the global population like no system ever did before.

## 3. Methodology

As presented in the previous chapters, several researchers have recently used deep learning techniques to identify depression using a multimodal approach. The majority of the researchers relied on the Extended-DAIC dataset [25] and DAIC WOZ [26], therefore, for possible comparison of our approach with others, we also decided to use the same dataset. These datasets include textual, audio, and video inputs from the interviews. Which data source provides the most successful symptom detection is a topic of continuing debate in the literature. Analysing data from the same dataset, Scherer et al. [27] focused only on analysing video inputs from the dataset. They found out that four behaviours linked with depression can be automatically recognized during the interview: angling of the head, eye gaze, duration and intensity of smiles, and self-touches. More precisely, participants with higher depression rates manifested significantly more downward head posture and eye gaze, smiled significantly less and with lower intensity and exhibited on average longer self-touches (hands rubbing, legs tapping and shaking). Despite the fact that nonverbal signs offer quite valuable information, they alone are not sufficient. Cummins et al. [5] mainly focused on the effects of depression and suicidality on common paralinguistic speech characteristics (prosody, source features, formant features, as well as spectral features). As stated by Huang et al. [18], numerous studies have demonstrated that depression can affect speech production in a variety of ways, including cognitive impairment, phonation and expression errors, articulatory incoordination, disturbances in muscle tension, psychomotoric delay, phoneme rates, as well as altered speech quality and prosody. Studies that use a multimodal approach tend to produce significantly better results than those that use one modality only [28,29]. However, modalities like audio and video raise a variety of ethical and data protection concerns that remain quite a big challenge that we still cannot tackle effectively [28]. Therefore, the data size and the efficiency of deep learning are the main reasons why most approaches are focused only on one of the modalities. For the same reason, in our approach, we also decided to focus only on textual-based inputs. To fuse more channels is our aim for further studies (Figure 1).

### 3.1. Metrics

For the evaluation of the results, the classification report from the sklearn-library is used on the test data. For a model comparison, we used metrics precision, recall and F1 score. These measures are the most commonly employed metrics to assess how well classification models perform, especially when it comes to binary classification, which has two classes: positive and negative. All three of these metrics are derived from the confusion matrix, a table that summarizes a classification algorithm's performance [30]. Precision refers to the proportion of true positive cases, or correctly diagnosed cases, out of all the cases that were identified as positive [31]. Precision can be particularly important for medical diagnosis as misdiagnosing a condition can have serious consequences for a patient's treatment and overall well-being. A high precision rate indicates that the diagnosis is likely to be correct, and therefore, the patient can receive appropriate treatment and care.

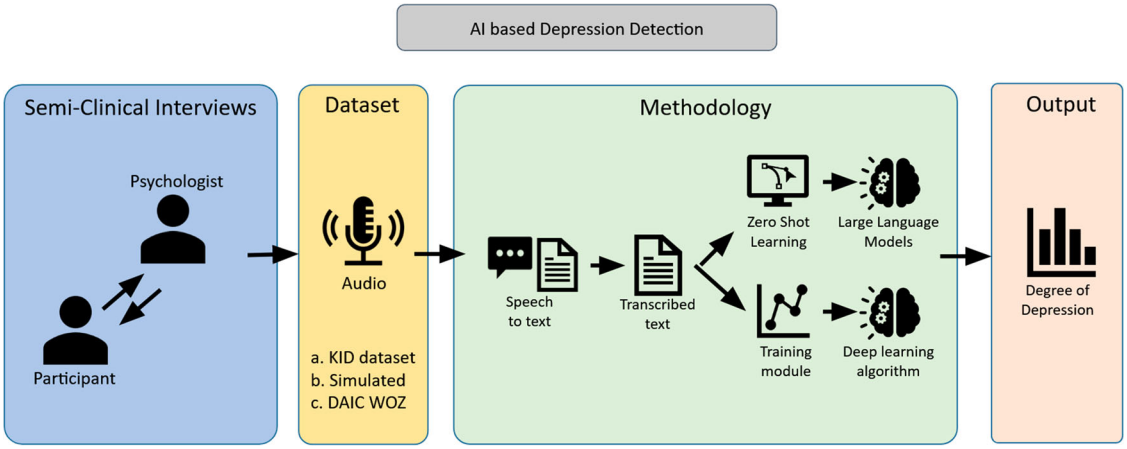



**Figure 1.** Overview of presented depression detection pipeline.

**Recall** also known as sensitivity, measures the proportion of actual positive cases that are correctly identified by the model [31]. A higher recall value means that the model is able to identify a larger proportion of positive cases, which is generally desirable in applications such as medical diagnosis, where the goal is to identify all cases of a disease, even if it means some false positives are identified.

**F1 score** is calculated as the harmonic mean of precision and recall. This means that the F1 score gives equal weight to both precision and recall [30]. A high F1 score indicates that a model has both high precision and high recall, which means it is able to correctly identify and classify cases accurately.

### 3.2. Datasets

The datasets used in this paper are all labelled with the PHQ-8 score as a ground truth value. The Patient Health Questionnaire-8 (PHQ-8) [32] is a widely used and validated self-report questionnaire designed to assess the severity of depressive symptoms in individuals. Consisting of eight items, the PHQ-8 is derived from the longer PHQ-9 by excluding the question related to suicidal ideation, which makes it a more suitable tool for research and settings where discussing suicidal thoughts may not be appropriate. Participants rate the frequency of experiencing each symptom over the past two weeks on a scale ranging from 0 (not at all) to 3 (nearly every day). The total score, ranging from 0 to 24, is calculated by summing the individual item scores, with higher scores indicating more severe depression. The PHQ-8 is valued for its brevity and simplicity, making it an efficient and reliable tool for screening depression in various populations and healthcare settings. Questionnaires and examination methods with more robust psychometric properties and research on how successful our work can support the diagnosis of medical specialists in the clinical practice should be further examined by further studies.

#### 3.2.1. DAIC-WOZ and Extended-DAIC datasets

As mentioned before, we used datasets, that are dominantly used in the research on this subject. The Distress Analysis Interview Corpus (DAIC) from the University of California – Institute for Creative Technologies (USCICT) offers two variants of datasets. The DAIC-WOZ [26] is based on the Wizard-of-Oz experiment and was developed for the research of anxiety, depression and post-traumatic stress disorder (PTSD) . The Extended-DAIC [25] is an extension of the DAIC-WOZ with a specialization on depression and PTSD. The data consists of transcribed text, audio and video recordings and was collected through interviews with the virtual interviewer Ellie, controlled by a human interviewer in another room, but participants were not aware of that. Both datasets consist of 456 recordings (196 in DAIC-WOZ and 263 in Extended-DAIC) with participants recruited as volunteers from the global population and veterans of US Army forces mostly diagnosed with depression, anxiety and/or PTSD.

#### 3.2.2. Simulated dataset

A novel approach to the test phase in our previous work [6] was that we used simulated data from standardized clinical interviews that were conducted especially for the purpose of this study. The interviews are conducted to measure variables from the PHQ-8 questionnaire. Facing specific ethical challenges regarding data protection when working with sensitive data, we wanted to test our models on the so-called 'simulated dataset'. The simulated data set is collected through interviews with psychology students who received comprehensive education and instructions on how depressed or non-depressed individuals would behave, communicate or react in the interview environment. Afterwards, mental health professionals conducted interviews with these students simulating the role they were assigned to. Their PHQ-8 scores as ground truth were given as a binary value, either depressed or non-depressed based on the role they got

to simulate in the interview. The so-called 'simulated data' in this study consists of transcribed interviews collected in this way. After collecting all the datasets for the testing phase (DAIC-WOZ, KID and Simulated dataset) we conducted a few steps to prepare them for the analysis. The first step was the transcription of the audio recording to textual data. For the transcription process, we utilized the speech-to-text model Whisper [33], developed by OpenAI. The whisper speech-to-text model is based on GPT (Generative Pre-trained Transformer) architecture. As it is developed specifically for this task, it has proved to be the best tool to perform this task. Additionally, it offers the possibility to translate the transcribed text into one of the preferred languages at the same time. We followed the same steps as Danner et al. [6] and also used Whisper to immediately translate the text from German to English.

#### 3.2.3. KID dataset

One of the novelties of this paper lies with the inclusion of another test dataset. In our previous work [6], we introduced a so-called simulated dataset that was intended to solve the ethical and data protection challenges we were facing. In the meanwhile, we successfully developed a study approach for obtaining real-world data in complete conformity with established criteria, all while upholding the highest ethical and data protection standards. The data protection procedure was approved by the Ethics Committee of the PFH Göttingen (*OS_18_200423*). From many volunteers drawn from the general population, we randomly selected 10 candidates with different levels of depressive symptoms for the interviews. Structured interviews with pre-trained students of psychology taking courses in clinical psychology. Prior to the interview, every participant filled out the PHQ-9 questionnaire and demographic information questionnaires. The used PHQ-9 scale is the version of the PHQ-8 questionnaire [32] with one more item regarding suicidal risk. As our fine-tuning was done on the DAIC-WOZ dataset where ground truth scores is PHQ-8, for the testing phase we also used the 8-item version without the item regarding suicidal intention. Interviews were conducted online, in German language, through an online conference platform that provides safe data storage and doesn't collect any personal information about the participants. During the interviews, only audio inputs were collected. Interviews were conducted following the GRID-HAMD-17 scale structure. GRID-HAMD-17 is an upgraded version of the Hamilton Depression Rating Scale (HDRS) [34] including the 17-item version of GRID-HAMD [35] scale. The GRID approach refers to rating the symptoms on a two-dimensional grid: its frequency and intensity. Interviews approximately lasted on average 15 minutes and were conducted online by the Private University of Applied Sciences Göttingen. In the further text, this dataset will be referred to as the KID dataset. The acronym originates from the German language standing for 'Künstliche Intelligenz in Depressionserkennung (Artificial Intelligence in Depression Detection).

### 3.3. BERT fine-tuning

The emergence of NLP transformer models, particularly Bidirectional Encoder Representations from Transformers (BERT), has marked a significant milestone in the field of AI. BERT, developed by researchers at Google, is a pre-trained language model that can be fine-tuned for a wide array of NLP tasks, such as sentiment analysis, machine translation, and question-answering systems. BERT's architecture leverages the transformer, which is an attention mechanism that learns contextual relationships between words or tokens in a text. Unlike traditional, unidirectional language models, BERT is designed to process input sequences bidirectionally, enabling it to capture both past and future context simultaneously. This bidirectional approach allows BERT to outperform its predecessors in numerous NLP benchmarks, setting new standards in the field. One of the key advantages of BERT is its ability to benefit from transfer learning, where the knowledge acquired from pre-training on vast amounts of data can be transferred to specific tasks with relatively small datasets. This characteristic not only reduces the need for extensive labelled data but also accelerates model training and improves overall performance.

In this work, we applied the following algorithms:

`BertTokenizer`: to split the text and tokenize sentences into subwords or wordpieces for the BERT model given a vocabulary generated from the *Wordpiece* algorithm.

`BertForSequenceClassification`: a BERT model transformer with a sequence classification and regression linear layer on top of the pooled output.

The data should be preprocessed for a more efficient classification process. First, the text from the interviewer was deleted and the rest of the transcript of the participants was saved in a string. The number of words in the transcribed text per participant ID in a string is more than 1,500 words. BERT itself can process a maximum of 512 tokens, which is the reason why the text of the participants is divided into fractions of 25 words. Then contractions like 'it's' and 'don't' that often occur in the English language will get written out with the Python library contractions. Then punctuation and the resulting double spacing, Zero values and numbers are removed. The classes for depressed and non-depressed participants are unevenly distributed as shown in Table 1. However, the class distribution

**Table 1.** Dataset class distribution.

| Dataset | depressed | non-depressed |
|---|---|---|
| DAIC-WOZ | 56 | 133 |
| Extended-DAIC | 66 | 209 |

should be even for the perception of a classification. To align the class distribution, an up-sampling of the minority class was performed. Additionally, we created more data by extracting the minority class and combining the second half of the current sentence and the first half of the next sentence of the extracted class, to get a new sentence with the same participant ID.

For the fine-tuning we started with a set of recommended hyperparameters from Delvin et al. [36]: AdamW optimizer with learning rate 3e-5, batch size of 32, and a number of epochs 4. With our model, we had a strong overfitting and to counteract that, we added a few regularizations. First, we tried to change the hidden and attention dropout rates after the work of El Anigri et al. [37]. On top of that, we also integrated a weight-decay over the AdamW optimizer. After the detailed process of fine-tuning and testing, we ended up with the parameters as presented in Table 2. These parameters could potentially still have more space for improvements in the future.

### 3.4. GPT-4 (API)

One of the differentiating factors of this work is the use of a brand-new model from OpenAI called GPT-4. This new model is bigger, multi-modal and hallucinates less. It surpasses the previous ChatGPT model by approximately 26% on the Multistate Bar Examination (MBE), even outperforming humans in this context [38]. It also comes out at the top in many other standardized tests and tasks such as code generation and reasoning. Concerning the whole preprocessing and data preparation process, we followed the same procedures as Danner et al. [6] to provide the same test environment and proper comparison of the results in these two studies

The Prompt was designed keeping in mind to let the model understand what the input text is and explicitly describe the form for the output. Table 3 depicts an example prompt used for the newer GPT-4 model. This newer model has a longer context length of 8192 tokens compared to only 4096 tokens. However, to keep up with the increasing demand, OpenAI aggressively the requests per minute rate as shown in the Table 4. This introduces additional complexity in handling these RateLimitErrors and extends the duration of task completion.

**Table 2.** BERT model parameters.

| Model | BERT-Base uncased |
|---|---|
| Environment | 12-layer, 768-hidden, 12-heads |
| Parameters | 110 M |
| Batch size | 16 |
| Length embedding | 27 |
| Epochs | 20 |
| Optimizer | AdamW |
| Learning rate | 3e-5 |
| Dropout hidden | 0.3 |
| Dropout attention | 0.3 |

**Table 3.** An example prompt used to predict the PHQ-8 scores of the participants using GPT-based models.

| | |
|---|---|
| **Prompt** | Give me one score (0 to 24 like the PHQ-8) for the whole interview. Only look at the Participant's answers. 0 is no depression, 24 is a severe depression. If the Score is greater or equal 10 the Participant is classified as 1 depressive. Only give me the score as an INTEGER WITHOUT EXPLANATION! [Interview] Interviewer: I'm going to ask you a battery of questions that all relate to the past seven weeks. Is it okay with you? Participant: Yes Interviewer: How was your mood during the past week? ⋯ . |
| **Output:** | 9 |

**Table 4.** Rate limits comparison between gpt-3.5 and gpt-4 models.

| Model | TPM[a] | RPM[b] |
|---|---|---|
| GPT-3.5-turbo | 160,000 | 5000 |
| GPT-4 | 10,000 | 200 |

[a]Tokens per minute.
[b]Requests per minute.

## 4. Results

In Table 5, we present the results and compare them to the other relevant studies on this topic. The mean precision, recall, and F1 scores are computed by initially determining these scores separately for both the positive (depressed) and negative (not depressed) classes, followed by averaging the results.

It is important to emphasize that some relevant studies were left excluded from this comparison because they used different or various datasets, fine-tuning approaches or presented metrics that are not comparable to others. Results have shown that GPT-4 outperforms other models on all datasets included in the comparison. Especially on the simulated dataset, where it has shown perfect performance. All other models have lower average recall values, suggesting that the

**Table 5.** Experimental results and comparison.

| Work | Precision | Recall | F1 score |
|---|---|---|---|
| DAIC-WOZ dataset $N = 42$ | | | |
| Villatoro-Tello (BERT)[39] | .59 | .59 | .59 |
| Senn et al. (BERT) [40] | | | .60 |
| Danner et al. (BERT) [6] | .63 | .66 | .64 |
| Danner et al. (GPT-3.5) [6] | .78 | .79 | .78 |
| Danner et al. (ChatGPT-4) [6] | .70 | .60 | .61 |
| Ours (GPT-4) | 0.81 | 0.70 | .71 |
| Simulated dataset $N = 5$ | | | |
| Danner et al. (BERT) [6] | .68 | .41 | .43 |
| Ours (GPT-4) | 1.00 | 1.00 | 1.00 |
| KID dataset $N = 10$ | | | |
| Ours (BERT) | .76 | .71 | .73 |
| Ours (GPT-3.5) | .83 | .70 | .73 |
| Ours (GPT-4) | .87 | .87 | .87 |

models misclassify the positive (depressed) class more often. GPT-4 model, however, consistently outperforms all other models across the board while achieving similar precision and recall, suggesting that it performs equally well for both classes. And, it could be concluded that this newer, bigger model has better generalization capacity because our KID dataset could not have been a part of its pre-training dataset. Therefore, to the best of our knowledge, our model outperforms the most recent state-of-the-art results when compared under the previously described comparison standards.

## 5. Discussion

At this point, it is important to emphasize once again that in our approaches, GPT models were not previously fine-tuned; instead, they were given specific prompts through API access. However, to this day, it remains unclear whether the PHQ-8 questionnaires or DAIC-corpus datasets were already known to the model during its training process as they were trained on a large amount of data available on various websites.

Currently, our work on this topic is still in its early stages. Therefore, one of the lacks of this study that we acknowledge is a relatively small dataset, but in the next steps of our approach, we are planning to include many more recordings as part of the KID dataset that we introduced in this paper. A larger KID dataset would allow us to analyse our results on a much deeper level, conduct appropriate statistical tests to compare the efficacy of the models, but also to include other variables when interpreting the results, such as demographic data, to test if the models are biased toward some specific subgroups.

While we recognize that domain adaptation in supervised learning models is often challenging, to set the tested models with more rigorous criteria, we decided to use a different dataset for testing our model from the one that was used for fine-tuning. In our previous work [6], we introduced a simulated dataset, which brings the potential to solve problems of data protection and safety, as well as to minimize and control the effects of various biases. Although our datasets were purposefully divergent, our results showed promising performance, indicating that currently dominant LLMs are flexible and can be used in a wide range of domains and test conditions. This feature, in our opinion, highlights the robustness and adaptability of the latest LLM, even in the face of changes between training and testing datasets, highlighting the potential of LLM development in the future.

Furthermore, our results demonstrate that GPT-based Large language models (LLMs) have the potential to detect depression reliably. Especially, the newer GPT-4 model performs exceptionally well compared to other models. This improved performance could be attributed to the fact that it is a bigger model with a larger context length and other undisclosed secret training recipes of OpenAI. This, along with the fact that it is trained on multiple modes of data, makes GPT-4 more generalizable, as evident by our experiments. However, the diagnostics interview data including transcripts, audio, and video is really sensitive. To this end, the use of commercially available open-source models would be really effective. Therefore, in future work, we would like to test open source models like GPT-J as they show comparable performance to that of GPT-based models and at the same time they have the potential to solve ethical and data protection issues when working with such a sensitive type of data.

As discussed in the related work section of the paper, there are some indications that other types of inputs, such as audio or video recordings could bring much more than just textual analysis, when analysing the interviews for depression screening. Therefore, in the next steps, this type of data could also be included to bring a new dimension to the analysis quality. Additionally, further studies are needed to explore which aspects of human-machine interaction in the context of depression detection interviews could have an impact on interaction quality. As stated in the introduction, our ultimate goal is to develop an automated AI-based depression detection model, where the interview process will be conducted by conversational AI.

In order to adjust our model to fulfil the needs of users and clinicians, acceptance and user experience-related studies should also be conducted in the next steps.

In the end, we would like to conclude with the statement that this model should not ever be used to make a final clinical diagnosis, but rather as a supporting tool for mental health experts to detect individuals during the early onset of depressive symptoms and recommend them further steps in order to receive appropriate support from the trained mental health experts.

## 6. Conclusion

In conclusion, this paper has demonstrated the impressive potential of the GPT-4 model in detecting depression from three various datasets. Our findings reveal that, compared to other methods, we could outperform other approaches in both accuracy and efficiency, thereby offering a robust and reliable means of identifying depression in individuals.

The superior performance of the GPT-4-based model underscores the importance of continued research into and development of large-scale language models for not only depression detection, but also for the detection of other mental health disorders and states. Given the increasing prevalence of mental health issues globally, developing automated tools capable of accurately detecting such conditions is crucial to providing timely and targeted support for those affected.

Looking ahead, we believe that fine-tuning large-scale language models could yield even more accurate and effective outcomes than the current ones. As such, we encourage further exploration of this promising field, with a particular emphasis on refining these models to become more sensitive to distinguish other disorders that often appear with comorbidity with depression.

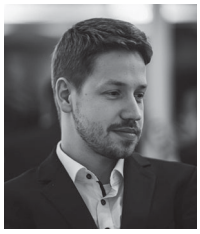
## Disclosure statement

No potential conflict of interest was reported by the author(s).

## Funding

This work was partially supported by the VwV Invest BW - Innovation II [grant number BW1_4056/03].

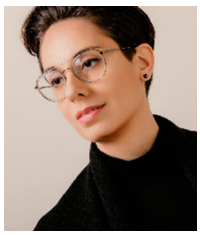
## Notes on contributors



*Bakir Hadzic* is a research assistant and PhD student at the Reutlingen University, Germany. His primary research interest revolves around the intersection of psychology and artificial intelligence, with a focus on acceptance, ethical and social aspects of AI in various application fields.

*Parvez Mohammed* is a research assistant and also a PhD student at Reutlingen University, Germany. He has obtained his Master degree from TU Dresden in the field of Computational Engineering. His research interests revolve around the application of AI in the field of Mental Health, Recruiting and Robotics.

*Michael Danner* is a research and teaching assistant at the Reutlingen University in Germany and PhD student at the University of Surrey, Guildford, United Kingdom. He is focused on AI and robotics research, specializing in robotic navigation, machine vision, and 3D reconstruction.



*Julia Ohse* is a research assistant and PhD student in the Department of Clinical Psychology at the Private University of Applied Sciences in Göttingen, Germany. Her research focus lies with advancing mental health screening through the application of Artificial Intelligence.

*Dr. Yihong Zhang* holds a role of Associate Professor at the College of Information Science and Technology. His research interests include image processing, classification, target tracking, and recognition. He is a Permanent Member of the Chinese Institute of Artificial Intelligence.

*Dr. Youssef Shiban* is a Professor of Clinical Psychology at the Private University of Applied Sciences in Göttingen, Germany, and a practicing psychotherapist. His research focuses on developing and optimizing prevention programs and treatment methods for anxiety disorders and depression, leveraging technical advancements such as Virtual Reality.

*Prof. Dr. rer. nat. Matthias Rätsch* is a professor for Artificial Intelligence and Interactive Mobile Robots at the Reutlingen University, and head of Visual Systems and Intelligent Robots – ViSiR Research Group. His research interests span various aspects of artificial intelligence, including image understanding, computer vision, deep learning, humanoid robots, bionic intelligence, human-robot collaboration, large language models, intelligent chatbots, and AI ethics.

## ORCID

*Bakir Hadzic* http://orcid.org/0009-0003-1197-7255
*Parvez Mohammed* http://orcid.org/0009-0001-7448-7857
*Michael Danner* http://orcid.org/0000-0002-8652-6905
*Julia Ohse* http://orcid.org/0009-0005-3344-4753
*Yihong Zhang* http://orcid.org/0000-0003-1261-1661
*Youssef Shiban* http://orcid.org/0000-0002-6281-0901
*Matthias Rätsch* http://orcid.org/0000-0002-8254-8293

## References

[1] World Health Organisation. World mental health report: Transforming mental health for all. World Health Organisation; 2022.
[2] Bech P. Rating scales in depression: limitations and pitfalls. In: Dialogues in clinical neuroscience, 2022.
[3] Radez J, Reardon T, Creswell C, et al. Why do children and adolescents (not) seek and access professional help for their mental health problems? A systematic review of quantitative and qualitative studies. Eur Child Adolesc Psychiatry. 2021;30:183–211. doi: 10.1007/s00787-019-01469-4
[4] Smith-East M, Neff DF. Mental health care access using geographic information systems: an integrative review. Issues Ment Health Nurs. 2020;41(2):113–121. doi: 10.1080/01612840.2019.1646363
[5] Cummins N, Scherer S, Krajewski J, et al. A review of depression and suicide risk assessment using speech analysis. Speech Commun. 2015;71:10–49. doi: 10.1016/j.specom.2015.03.004
[6] Danner M, Hadzic B, Gerhardt S, et al. Advancing mental health diagnostics: Gpt-based method for depression detection. In: Proceedings Title; 62nd Annual Conference of the Society of Instrument and Control Engineers of Japan (SICE); Tsu City, Japan; 2023; p. 1290–1296.
[7] Calvo R, D'Mello S, Gratch J, et al. Cyberpsychology and affective computing. In: The Oxford handbook of affective computing. Oxford University Press, Jan. 2015.
[8] Scherer KR. What are emotions? and how can they be measured?. Soc Sci Inf. 2005 Dec;44(4):695–729. doi: 10.1177/0539018405058216

[9] Alghowinem S, Goecke R, Epps J, et al. Cross-cultural depression recognition from vocal biomarkers. In: Interspeech 2016; Sep. ISCA; 2016.
[10] Park J, Moon N. Design and implementation of attention depression detection model based on multimodal analysis. Sustainability. 2022 Mar;14(6):3569. doi: 10.3390/su14063569
[11] Uslu I. Deep Learning im Mental Health Kontext [master's thesis]. Reutlingen University, 2023.
[12] Raffel C, Shazeer N, Roberts A, et al. Exploring the limits of transfer learning with a unified text-to-text transformer. In: 21.1, 2020.
[13] Cheng JC, Chen ALP. Multimodal time-aware attention networks for depression detection. J Intell Inf Syst. 2022 Apr;59(2):319–339. doi: 10.1007/s10844-022-00704-w
[14] Toto E, Tlachac M, Rundensteiner EA. AudiBERT. In: Proceedings of the 30th ACM International Conference on Information & Knowledge Management; Oct. ACM; 2021.
[15] Sharma A, Sharma K, Kumar A. Real-time emotional health detection using fine-tuned transfer networks with multimodal fusion. Neural Comput Appl. 2023 Jan;35(31):22935–22948.
[16] Aloshban N, Esposito A, Vinciarelli A. What you say or how you say it? Depression detection through joint modeling of linguistic and acoustic aspects of speech. Cognit Comput. 2022;14(5):1585–1598. doi: 10.1007/s12559-020-09808-3
[17] Davis S, Mermelstein P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Trans Acoust. 1980;28(4):357–366. doi: 10.1109/TASSP.1980.1163420
[18] Huang Z, Epps J, Joachim D. Investigation of speech landmark patterns for depression detection. IEEE Trans Affect Comput. 2022;13(2):666–679. doi: 10.1109/TAFFC.2019.2944380
[19] McGinnis EW, Anderau SP, Hruschak J, et al. Giving voice to vulnerable children: machine learning analysis of speech detects anxiety and depression in early childhood. IEEE J Biomed Health Inform. 2019;23(6):2294–2301. doi: 10.1109/JBHI.6221020
[20] Bailey A, Plumbley MD. Gender bias in depression detection using audio features. In: 29th European Signal Processing Conference, EUSIPCO 2021, Dublin, Ireland, August 23-27, 2021. IEEE; 2021. p. 596–600.
[21] OpenAI. GPT-4 technical report. In: CoRR abs/2303.08774; 2023.
[22] Thoppilan R, Freitas DD, Hall J, et al. Lamda: Language models for dialog applications. In: CoRR abs/2201.08239; 2022.
[23] Sun Y, Wang S, Feng S, et al. ERNIE 3.0: Large-scale knowledge enhanced pre-training for language understanding and generation. In: CoRR abs/2107.02137; 2021.
[24] Touvron H, Lavril T, Izacard G, et al. Llama: Open and efficient foundation language models. In: CoRR abs/2302.13971; 2023.
[25] DeVault D, Artstein R, Benn G, et al. Simsensei kiosk: A virtual human interviewer for healthcare decision support. In: Proceedings of the 2014 international conference on Autonomous agents and multi-agent systems; 2014; p. 1061–1068.
[26] Gratch J, Artstein R, Lucas G, et al. The distress analysis interview corpus of human and computer interviews. Tech. rep. University of Southern California Los Angeles, 2014.
[27] Scherer S, Stratou G, Lucas G, et al. Automatic audiovisual behavior descriptors for psychological disorder analysis. Image Vis Comput. 2014;32(10):648–658. doi: 10.1016/j.imavis.2014.06.001
[28] Lopez-Otero P, Docio-Fernandez L. Analysis of gender and identity issues in depression detection on de-identified speech. Comput Speech Lang. 2021;65:101118. doi: 10.1016/j.csl.2020.101118
[29] Qureshi SA, Saha S, Hasanuzzaman M, et al. Multitask representation learning for multimodal estimation of depression level. IEEE Intell Syst. 2019;34(5):45–52. doi: 10.1109/MIS.2019.2925204
[30] Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over f1 score and accuracy in binary classification evaluation. BMC Genomics. 2020;21(1):1–13. doi: 10.1186/s12864-019-6413-7
[31] Yacouby R, Axman D. Probabilistic extension of precision, recall, and f1 score for more thorough evaluation of classification models. In: Proceedings of the first workshop on evaluation and comparison of NLP systems; 2020; p. 79–91.
[32] Kroenke K, Strine TW, Spitzer RL, et al. The PHQ-8 as a measure of current depression in the general population. J Affect Disord. 2009;114(1-3):163–173. doi: 10.1016/j.jad.2008.06.026
[33] Radford A, Kim JW, Xu T, et al. Robust speech recognition via large-scale weak supervision. In International Conference on Machine Learning; 2023; p. 28492–28518. PMLR.
[34] Hamilton M. A rating scale for depression. J Neurol Neurosurg Psychiatr. 1960;23(1):56–62. doi: 10.1136/jnnp.23.1.56
[35] Itai A, Papadimitriou CH, Szwarcfiter JL. Hamilton paths in grid graphs. SIAM J Comput. 1982;11(4):676–686. doi: 10.1137/0211056
[36] Devlin J, Chang MW, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding; 2018.
[37] El Anigri S, Himmi MM, Mahmoudi A. How bert's dropout fine-tuning affects text classification?. In: Business Intelligence; 2021; p. 130–139.
[38] Katz DM, Bommarito MJ, Gao S, et al. Gpt-4 passes the bar exam. 2023 March 15.
[39] Villatoro-Tello E, Ramirez-de-la Rosa G, Gática-Pérez D, et al. Approximating the mental lexicon from clinical interviews as a support tool for depression detection. In: Proceedings of the 2021 International Conference on Multimodal Interaction; 2021; p. 557–566.
[40] Senn S, Tlachac M, Flores R, et al. Ensembles of bert for depression classification. In: 2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC); 2022; p. 4691–4694.