



# What You Say or How You Say It? Depression Detection Through Joint Modeling of Linguistic and Acoustic Aspects of Speech

Nujud Alosbhan<sup>1,2</sup> · Anna Esposito<sup>3,4</sup> · Alessandro Vinciarelli<sup>2</sup>

Received: 30 July 2020 / Accepted: 7 December 2020 / Published online: 24 February 2021  
© The Author(s) 2021

## Abstract

Depression is one of the most common mental health issues. (It affects more than 4% of the world’s population, according to recent estimates.) This article shows that the joint analysis of linguistic and acoustic aspects of speech allows one to discriminate between depressed and nondepressed speakers with an accuracy above 80%. The approach used in the work is based on networks designed for sequence modeling (bidirectional Long-Short Term Memory networks) and multimodal analysis methodologies (late fusion, joint representation and gated multimodal units). The experiments were performed over a corpus of 59 interviews (roughly 4 hours of material) involving 29 individuals diagnosed with depression and 30 control participants. In addition to an accuracy of 80%, the results show that multimodal approaches perform better than unimodal ones owing to people’s tendency to manifest their condition through one modality only, a source of diversity across unimodal approaches. In addition, the experiments show that it is possible to measure the “confidence” of the approach and automatically identify a subset of the test data in which the performance is above a predefined threshold. It is possible to effectively detect depression by using unobtrusive and inexpensive technologies based on the automatic analysis of speech and language.

**Keywords** Depression · Computational paralinguistics · Word embedding · Bidirectional long-short term memory networks · Joint representation · Late fusion

## Introduction

Clinical depression is a pathology resulting from two main processes, namely “an *increase* in negative emotions and feelings and a *reduction* in positive emotions and feelings” [1]. As a consequence, patients experience extended periods of “depressed mood, loss of interest or pleasure, decreased energy, feelings of guilt or

low self-worth, disturbed sleep or appetite, and poor concentration” [2]. In terms of diffusion and impact on the quality of life of people, the World Health Organization estimates that depression affects 4.4% of the world’s population (approximately 300 million people in 2015), and it is the most common cause of disability (it accounts for 7.5% of all years lived with disability in 2015) and suicide (close to 800,000 cases per year) [3]. Nevertheless, such figures might be an underestimate because multiple factors, including stigma and lack of available services, tend to discourage depressed individuals from seeking treatment, resulting in many undetected cases, especially in regard to “[...] developing countries, older cohorts, men, and cases with earlier ages of onset” [4].

The literature shows that artificial intelligence (AI) can contribute to addressing these problems through automatic technologies for depression detection (see [Survey of Previous Work](#) for a survey). In particular, the clinical practice of psychiatrists, based on the observation of patients’ behavior, lends itself to the application of methodologies for the inference of psychological phenomena from machine-detectable behavioral cues (e.g., social signal

---

✉ Alessandro Vinciarelli  
vincia@dcs.gla.ac.uk; Alessandro.Vinciarelli@glasgow.ac.uk

Nujud Alosbhan  
n.aloshban.1@research.gla.ac.uk

Anna Esposito  
Anna.Esposito@unicampania.it

<sup>1</sup> Imam Muhammad ibn Saud Islamic University in Riyadh, Riyadh, Saudi Arabia

<sup>2</sup> University of Glasgow, Glasgow, Scotland, UK

<sup>3</sup> Università degli Studi della Campania “L. Vanvitelli”, Caserta, Italy

<sup>4</sup> IIASS, Vietry sul Mare, Italy

processing [5], affective computing [6] or computational paralinguistics [7]). For this reason, this article proposes a multimodal approach designed to detect depression based on linguistic and acoustic aspects of speech. In particular, the article uses network architectures combining speech signals and their transcriptions through *joint representations* [8] and *gated multimodal units* [9] that take into account both *what* people say and *how* they say it.

The experiments were performed over a corpus of 59 interviews, approximately 4 hours of material, collected in three mental health centers in Southern Italy. The total number of experiment participants was 59, including 29 persons diagnosed with depression by a psychiatrist and 30 who had never experienced mental health issues. The results show that the multimodal approaches performed better, to a statistically significant extent, than the unimodal ones. In particular, the best multimodal approach achieved an accuracy of 84.7% (F1 measure of 82.3%), indicating that it correctly distinguishes between depressed and nondepressed speakers approximately 4 times out of 5.

During the experiments, the approaches were applied to *clauses*, i.e., to manually extracted linguistic units that include a noun, a verb and a complement. Given that the average number of clauses per participant is 114, this allows, for every person, a large number of clause-level decisions. These can then be aggregated through a majority vote, which, despite being accurate less than 70% of the time, results in an accuracy of up to 84.7% in determining whether a person is depressed. This is important because it shows that it is possible to deal effectively with the limited amount of available data, a problem that is inherent to depression detection due to ethical and practical concerns in recruiting depression patients.

In addition to the above, the experiments showed that, unlike in other studies [10, 11], acoustic aspects of speech appear to be more effective than linguistic ones in conveying depression-relevant information (despite the clauses having been transcribed manually). One possible explanation is that approaches based on language have difficulty dealing with short linguistic units such as clauses (for which the average number of words is 3.9). However, another possible reason is that *paralinguistics* (how things are said) might be a more honest cue than *lexical choice* (what people say), at least in regard to the features used in this work. This is in line with the observation of social psychology that nonverbal behavior, because it is displayed outside conscious awareness, tends to convey more reliable information about the inner state of an individual [12].

To the best of our knowledge, this is one of the first depression detection works involving Italian speakers. This is important because it shows that depression detection technologies can be effective not only for English speakers, the most common cases in the literature, but also for people

who belong to different cultures. Furthermore, in contrast to other studies (see [Survey of Previous Work](#)), the distinction between depressed and nondepressed participants has usually been made by psychiatrists and not through the administration of self-assessment questionnaires. This is an advantage because it increases the chances of the data being representative of the actual difference between depressed and nondepressed speakers. In other words, it ensures that the problem addressed in the work is depression detection and not an inference of the self-assessment scores. This is important because self-assessment questionnaires are subject to multiple biases [13]; furthermore, the data show that they can be filled out inconsistently, especially by people affected by depression (see [The Data](#)).

In addition to the above, the experiments showed that roughly one-third of the participants tended to manifest their condition through one modality or another. In particular, depressed participants tended to do so through the way they spoke (acoustic aspects), while control participants tended to do so through the words they used (linguistic aspects). To the best of our knowledge, such an observation has not previously been made in the literature, and it is important because it explains the *diversity* of the unimodal approaches (the tendency to make different mistakes over different samples) [14], which is probably the main reason for the effectiveness of the multimodal combination.

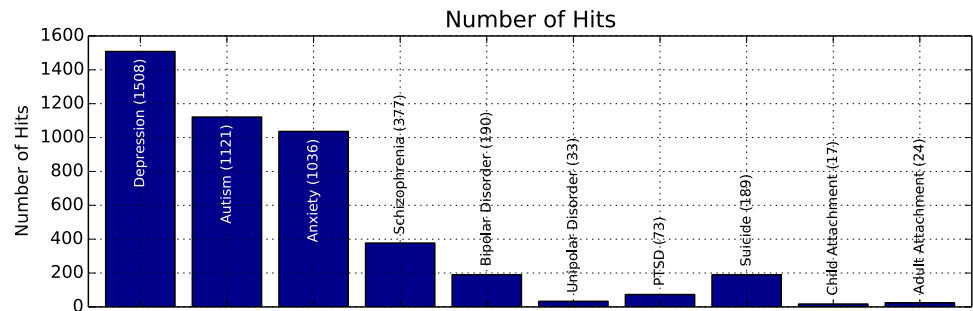
The rest of this article is organized as follows: [Survey of Previous Work](#) provides a survey of the literature, [The Data](#) describes the data used in the experiments, [The Approach](#) describes the proposed approach, [Experiments and Results](#) reports on the experiments and results, and [Discussion and Conclusions](#) draws some conclusions.

## Survey of Previous Work

[Introduction](#) shows that depression has a major impact not only on the lives of patients but also on society as a whole. [Figure 1](#) shows that when the query “*depression psychiatry*” is submitted, *IEEEExplore* returns more hits than for any other mental health issue. In fact, depression has been the subject of at least four benchmarking campaigns organized in the last decade, including two based on a corpus that shows 292 people performing a human-computer interaction task [15, 16] and two based on a corpus of more than 200 individuals interacting with an artificial agent [17, 18]. In all cases, the task is to infer the scores resulting from the administration of self-assessment questionnaires such as the Beck Depression Inventory II (BDI-II) [37] or different versions of the *Patient Health Questionnaire* (PHQ) [19].

While the works mentioned so far have focused on the inference of self-assessment scores, others have addressed the problem of professional psychiatrists detecting people

**Fig. 1** The chart shows the number of hits returned when submitting queries related to mental health issues to IEEEExplore (<https://ieeexplore.ieee.org/Xplore/home.jsp>). The queries have been submitted with the constraint of returning material published after 2009



with depression (as this article does) [20–24]. Such a task is performed with accuracy up to 90% of the time in [21] with the use of electro-encephalograms (EEG); with 88% accuracy through the multimodal combination of paralinguistics, head pose and gaze in an approach presented in [20]; and with an F1 measure up to 80% by analyzing body movement in combination with head pose and facial expressions [23]. While these results are not being comparable (they were not obtained with the same data), they suggest that it is possible to replicate the judgment of professional psychiatrists approximately 4 times out of 5.

In regard to the modalities used in this article (linguistic and acoustic aspects of speech), several works have proposed experiments to investigate specific aspects of depression. In [25], the focus is on the use of short utterances collected through mobile phones (a setting typical of counseling services accessible through the phone). The results show that it is possible to detect people above the PHQ-9 threshold corresponding to depression with accuracy up to 72% of the time. The experiments proposed in [26] address the problem of adolescent voices, which, not being fully formed, are more challenging to process automatically. The results show that energy, accounting for how loudly people speak, is the best depression marker, especially when measured with the Teager operator [7]. Similarly, the results presented in [27] indicate that the main difference between depressed and nondepressed speakers is phonetic variability, with depressed people tending to be less variable.

In addition to the above, several works have addressed the problem of combining speech and its transcription (as this work does). While some works have suggested, based on experimental evidence, that linguistic and paralinguistic aspects of speech should always be modeled jointly [28], others have shown that this is not necessarily the case and that better results can be achieved, e.g., by using solely speech transcriptions [11]. Furthermore, other works have suggested that the multimodal combination of speech and its transcription improves over individual modalities only when taking into account when a sentence is uttered during an interaction [29] or by using models that include attention gates capable of identifying, for every sample, the modality most likely to produce the best results [10]. In other words,

it is unclear whether depression-relevant information is transmitted more effectively by linguistic or acoustic aspects of speech.

Overall, the state-of-the-art summary presented in this section suggests that no form of behavioral evidence (speech, facial expressions, gestures, etc.) clearly outperforms the others. Furthermore, the use of similar approaches (e.g., the joint modeling of linguistic and acoustic aspects of speech [10, 28, 29]) over different data does not necessarily lead to the same conclusions about how effective the use of a certain modality is with respect to the others. One possible reason for such a state of affairs is that several works address the problem of identifying people diagnosed with depression not on the basis of a doctor’s evaluation but rather by inference from self-assessment scores. These are affected by different biases (see [The Data](#)) and, therefore, can lead to ambiguous results. Furthermore, depression is a complex phenomenon that involves a wide spectrum of factors (e.g., physiology, socioeconomic status, age and gender [1]) that result in individual differences in the way people manifest the pathology.

## The Data

The experiments were performed over a corpus of 59 interviews recorded in three mental health centers in Southern Italy. Every interview involved a different participant, but the protocol was always the same. In particular, the interviewers always posed the same questions (e.g., “*What did you do during the last week-end?*”) and always in the same order. Out of the 59 participants, 29 had been diagnosed with depression by a professional psychiatrist, while the remaining 30, referred to as *control* participants, had never experienced mental health issues. The interviewers were instructed to speak as little as possible, and on average, they spoke for 10.0% of the interview duration: 5.1% with the depressed participants and 14.7% with the control participants. The difference is statistically significant ( $p < 10^{-5}$  according to a two-tailed  $t$ -test), and one possible explanation is that the control participants tended to involve the interviewers in the interaction, while the depressed ones simply tended to answer the questions.

**Table 1** The table shows the demographic information available for the participants. According to a *t*-test, there was no difference between the depressed and control participants in terms of age. Simi-

larly, according to a  $\chi^2$  test, the distribution of gender and education level was the same for both groups

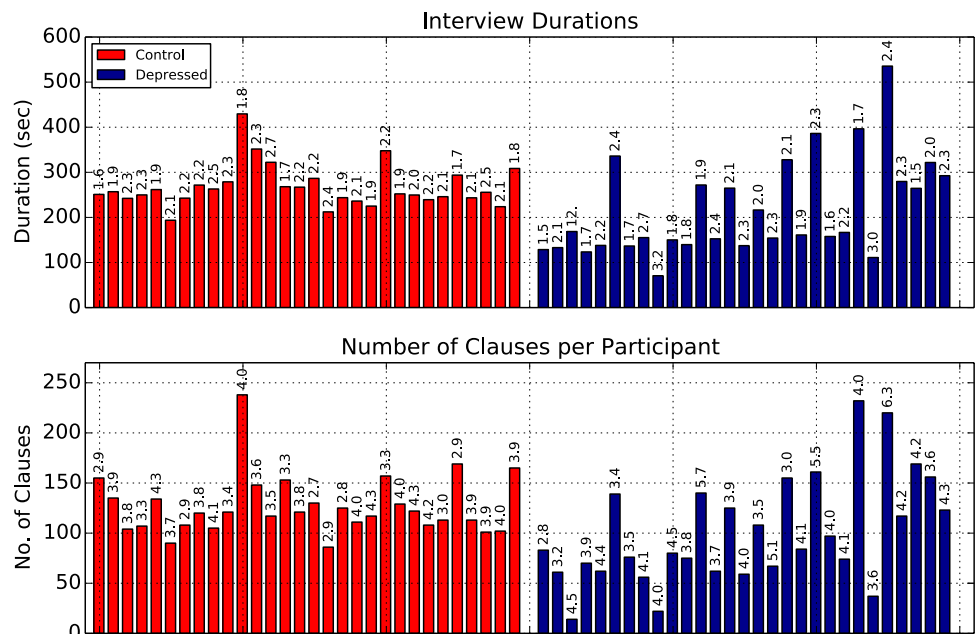
	F	M	Avg. Age	Age Range	Primary	Superior
Depressed	21	8	45.7	23-69	16	13
Control	21	9	44.0	23-68	12	18
Total	42	18	44.4	23-69	28	31

Table 1 provides demographic information. The gender distribution was the same for both the depressed and control groups, with 2.47 times more females than males. This is in line with the observation that despite cultural and national differences [30], women tend to develop depression roughly two times more frequently than men [31]. In terms of age, the range is roughly the same, and according to a two-tailed *t*-test, there is no statistically significant difference between the average ages (45.7 for the depressed group and 44.0 for the control group). The motivation behind the choice of the range was that depression tends to be less frequent among children [32], adolescents [33] and people older than 65 [34]. In this respect, the experiment participants should be representative of the population most likely to develop depression. Finally, the table reports the distribution across the education levels of the Italian system, namely, *Primary* (up to 8 years of education) and *Superior* (between 13 and 18 years of education). According to a two-tailed  $\chi^2$  test, the difference between the two distributions is not statistically significant. Overall, the two groups differed in terms of mental health condition (depressed or control) but

not in terms of other factors (gender, age and education). This should ensure that the approach proposed in this work detects depression and not other factors that might result in linguistic or acoustic differences in speech.

The upper chart of Fig. 2 shows how durations were distributed across the participants. On average, every interview lasted 242.2 seconds, but there was a statistically significant difference ( $p < 0.05$  according to a one-tailed *t*-test) when the depressed and control participants were considered separately (the averages were 216.5 and 267.1 seconds, respectively). Every interview was manually transcribed and segmented into clauses, i.e., basic linguistic units that include a noun, a verb and a complement. The clauses were the analysis unit of the experiments, meaning that they were analyzed and recognized individually before a participant was classified as depressed or control (see [Experiments and Results](#) for more detail). For this reason, the lower chart of Fig. 2 shows the distribution of clauses and the average number of words they include. Overall, the average number of clauses is 114.0, but it is 100.8 and 126.8 for the depressed and control participants, respectively (the difference is statistically significant, with  $p < 0.05$  according

**Fig. 2** The upper chart shows the interview durations for all participants, and the number at the top of each bar is the average duration (in seconds) of each clause. The lower chart shows the number of clauses for each participant, and the number at the top of each bar is the average number of tokens per clause (tokens are sequences of characters enclosed between two consecutive blank spaces and typically correspond to words). In both charts, the depressed and control participants are shown separately



**Table 2** The table shows the distribution of scores across the four conventional ranges used to interpret the Beck Depression Inventory II scores, namely, *minimal* (0-13), *mild* (14-19), *moderate* (20-28), and *severe* (29-63)

Condition	Minimal	Mild	Moderate	Severe
Depressed	7	2	11	6
Control	24	2	1	2
Total	31	4	12	8

to a one-tailed *t*-test). However, there is no statistically significant difference in terms of the average number of words per participant (429.7 and 463.9 for the depressed and control participants, respectively), suggesting that the depressed participants tended to use more words per clause. The differences in duration and number of clauses are compatible with previous observations showing that people affected by depression tend to display lower involvement in conversations [35, 36].

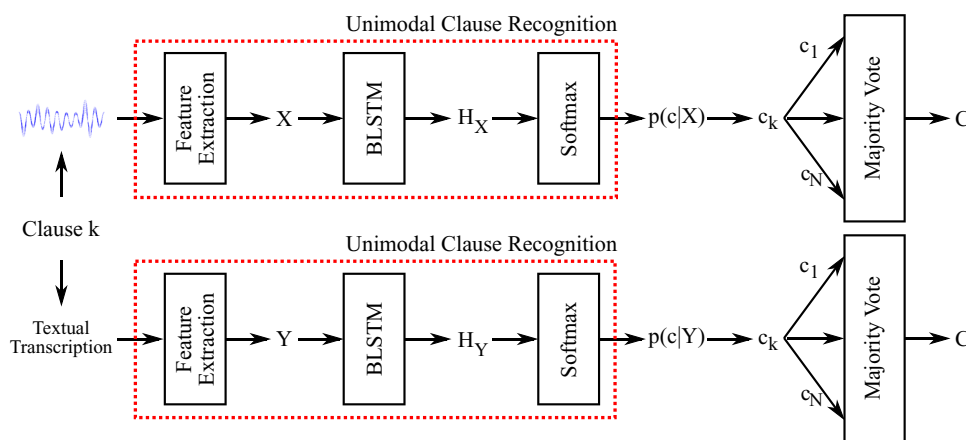
Out of the 59 participants, 55 filled out the *Beck Depression Inventory II* (BDI-II) [37], one of the self-assessment questionnaires most commonly used to support the diagnosis of depression. The result of the questionnaire is a score that, on average, is proportional to the severity of the depression condition. Table 2 shows the distribution of the scores across the four conventional ranges used to interpret the BDI-II scores, namely, *minimal* (0-13), *mild* (14-19), *moderate* (20-28) and *severe* (29-63). The data show that on average, the scores accounted for the actual condition of the participants (the average scores were 21.7 and 9.7 for the depressed and control participants, respectively). However, roughly one-third of the participants diagnosed with depression had scores that fell in the minimal and

mild ranges, which are not considered pathological. This suggests that the BDI-II scores, at least in the data used for this work, cannot be considered fully reliable, especially in regard to depression patients. One possible explanation is that self-assessment questionnaires have been shown to be sensitive to multiple biases, and “[...] accuracy is not the only motive shaping self-perceptions [...] the other powerful motives are consistency seeking, self-enhancement, and self-presentation” [13]. In other words, the data shown in Table 2 suggest that several depressed participants were unable to fill out the questionnaire or possibly tried to conceal their condition, perhaps to avoid the stigma associated with mental health issues.

### The Approach

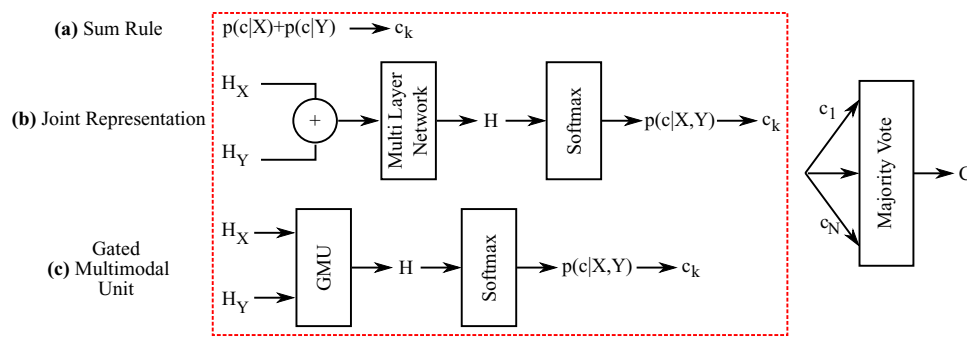
Figure 3 shows the unimodal recognition approach used in the experiments. The feature extraction maps the clauses into sequences of feature vectors that are then fed to a bidirectional long short-term memory network (BLSTM) [38]. The BLSTM outputs a representation that is provided as input to a softmax layer that estimates the posterior probabilities of the two possible classes (*control* and *depression*). A clause is then assigned to the class with the highest a posteriori probability, and the classification outcomes corresponding to the *N* clauses uttered by a particular individual are aggregated through a majority vote. In other words, an individual is assigned to the class her or his clauses are most frequently assigned to.

Figure 3 shows the different strategies of multimodal recognition approaches. In particular, the unimodal representations output by the BLSTMs ( $H_X$  and  $H_Y$  in Fig. 4)



**Fig. 3** The figure shows the unimodal recognition approach. Speech signal and textual transcription corresponding to every clause *k* are converted into sequences of feature vectors (*X* and *Y*, respectively) that are fed to a BLSTM followed by a softmax layer. The output of the latter can be thought of as the a posteriori probability distribu-

tion of the classes (a clause is assigned to the class with the highest a posteriori probability). The classification outcomes of the individual clauses are aggregated through a majority vote (a participant is assigned to the class her or his clauses are most frequently assigned to)



**Fig. 4** The figure shows the three strategies for the multimodal combination of linguistic and acoustic aspects of speech. The sum rule (or late fusion) makes use of the unimodal posteriors as a criterion for assigning a clause to a given class. Joint representation “fuses” the unimodal representations (see [Unimodal Recognition](#)) through a

are combined through the *joint representation* (JR) [38] and *gated multimodal unit* (GMU) [9] multimodal recognition approaches. In addition, the output of the unimodal classifiers serves as input to the *sum rule* [39] multimodal approach. The use of different combinations of approaches ensures that the conclusions of this work result from actual properties of the data and not from the use of a particular methodology.

The rest of this section provides more details about the feature extraction process (see [Feature Extraction](#)), unimodal recognition (see [Unimodal Recognition](#)) and multimodal recognition (see [Multimodal Recognition](#)).

## Feature Extraction

For every clause, the data correspond to a speech signal and its manual transcription. Therefore, the feature extraction step processes speech and text separately and produces two different sequences of vectors, a format suitable for the use of BLSTMs (see [Unimodal Recognition](#)).

In the case of speech, the signal is segmented into 25 ms long analysis windows that start at regular time steps of 10 ms and span the entire clause (two consecutive windows overlap by 15 ms). The values of both window length and step are standard in the literature, and no other values have been tested. After the segmentation, the signal interval enclosed in every window is mapped into a feature vector where the components are the first 39 *mel frequency cepstral coefficients* (MFCC) [7]. Such a representation, based on the physiology of hearing, is widely applied in the literature and accounts mainly for energy (how loudly someone speaks) and phonetic content (what sounds someone utters). The main motivation behind its use is that it has been effective in a wide spectrum of approaches aimed at inferring social and psychological phenomena from speech, including, e.g., emotions [40], personality [41] and depression (see [Survey of Previous Work](#)). As a result, every clause is converted

into a sequence of vectors  $X = (x_1, x_2, \dots, x_T)$ , where  $x_i$  is the vector extracted from the  $i^{\text{th}}$  window and  $T$  is the total number of vectors.

4-layer network that takes as input the concatenation of  $H_X$  and  $H_Y$ . Finally, the gated multimodal unit weights the unimodal representations according to how likely they are to lead to the correct classification outcome

In the case of the clause transcriptions, the text is converted into a sequence  $Y = (y_1, y_2, \dots, y_N)$  of vectors, one for every word in the clause, through the application of *Wikipedia2Vec* [42], an extension of *Word2vec* that learns word embedding [43]. The key idea of such a methodology is to use a shallow network (only one hidden layer) to map every word in a text to the word that follows it [44]. If  $V$  is the dictionary (list of unique words taken into account),  $|V|$  its size and  $h$  the number of hidden neurons in the shallow network (with  $h \ll |V|$ ), the training results in a matrix  $W$  of weights connecting the input layer to the hidden layer ( $h = 100$  in the experiments of this work). In such a matrix, the element  $w_{ij}$  can be thought of as the  $j^{\text{th}}$  component of the vector describing the  $i^{\text{th}}$  word of the dictionary. During the training, input and output words are represented with a *one-hot* vector of dimension  $|V|$  (the  $k^{\text{th}}$  word of the dictionary is represented with a vector where all elements are set to 0 except element  $k$ , which is set to 1). The experiments of this work are based on a version of *Wikipedia2vec* pretrained on a corpus of Italian texts, “*itwiki*”, including Wikipedia articles written in Italian.

While showing that *Wikipedia2vec* is effective in a wider range of applications, the literature proposes more sophisticated word-embedding methodologies that take into account the context and, in particular, represent the same word in different ways depending on the context in which the word appears (e.g., *bidirectional encoder representations from transformers* (BERT) [45]). However, in the experiments of this work, such methodologies did not result in any improvements. The main probable reason is that the clauses are short (the average length is 3.9 words), and therefore, the context might not carry sufficient information. For this reason, the experiments rely on the sole use of *Wikipedia2vec*.

## Unimodal Recognition

The feature extraction step converts speech signals and the associated transcriptions into sequences of feature vectors (see previous section). Such a representation is suitable for the application of BLSTMs [38], models allowing the classification of sequential data. The main advantage of BLSTMs is that they take into account both past and future contexts of every element of an input sequence because of the presence of two hidden layers, not connected to each other and connected to the output layer, that have the same architecture as the hidden layer of a standard LSTM. One of the hidden layers processes the input sequences from the first element to the last, and, conversely, the other one processes it from the last sample to the first.

Unimodal recognition takes place by training two BLSTMs separately, one per modality, with a *softmax* output layer. In such a way, the network outcome corresponds to the *a posteriori* probability of sequences  $X$  and  $Y$  (see previous section) belonging to classes *depressed* or *control*. A sequence is then assigned to the class corresponding to the highest *a posteriori* probability.

## Multimodal Recognition

Figure 4 shows the three strategies for the multimodal combination of linguistic and acoustic aspects of speech. Given that both unimodal clause recognition approaches output *a-posteriori* probabilities, the most straightforward approach to multimodal recognition is the application of the *sum rule* (see Fig. 4a), probably the most commonly applied technique for the combination of multiple classifiers [39]:

$$\hat{c} = \arg \max_{c \in \mathcal{C}} \{p(c|X) + p(c|Y)\} \quad (1)$$

where  $\hat{c}$  is the class assigned to a clause,  $\mathcal{C}$  is the set of all possible classes (*control* and *depression* in the experiments of this work),  $X$  is the sequence extracted from speech and  $Y$  is the sequence extracted from text (see above). Originally developed for combining different classifiers fed with the same input, the sum rule is often referred to as *late fusion* when used to combine the output of classifiers working on different modalities (as in the case of this work). While based on the unrealistic assumption that  $X$  and  $Y$  are statistically independent given the class, the approach has been shown to be effective in a wide range of problems.

The other typical approach for multimodal recognition is *early fusion*, i.e., the concatenation of feature vectors extracted at the same moment from multiple modalities. The problem is that in the experiments of this work, there is a significant difference in the rate at which the vectors are extracted from the data. In the case of speech signals, one vector is extracted every 10 *ms*, thus resulting in a rate of

100 *Hz*, while in the case of texts, there is one vector per word, thus resulting in a rate of roughly 2 *Hz* (the average number of words per second). In such a situation, the application of the early fusion requires downsampling the sequence where the rate is greater and, as a consequence, discarding information. However, it is possible to avoid such a problem by obtaining a *joint representation* (JR) through the approach depicted in Fig. 4b. The unimodal representations are extracted from the last hidden state of BLSTM ( $H$ ) from speech and its transcription. The key idea behind such an approach is to use the output of the hidden layer as a representation of the entire sequence. These representations are then concatenated and fed to a 4-layer network that is expected to output a representation that better discriminates between the depressed and control participants.

Finally, a third multimodal approach corresponds to feeding the unimodal representations to a *gated multimodal unit* (GMU) [9], as shown in Fig. 4c. The GMU learns to weight the representations of the two modalities and, in particular, to increase the weight of the modality that appears to carry depression-relevant information. In the experiments of this work, the weights are scalar parameters  $w_i$  that can be thought of as a measure of the impact that modality  $i$  has on the classification outcome.

## Experiments and Results

The experiments were performed according to a  $k$ -fold experimental design. The participants were randomly split into  $k = 5$  disjoint subsets, and the clauses uttered by all participants in the  $k - 1$  groups were used as the training set. Correspondingly, the clauses uttered by the participants in the left-out subset were used for the test set. The process was repeated  $k$  times, and at each repetition, a different subset was left out. This made it possible to perform experiments over the whole corpus at disposition while still keeping separate training and test sets. Another advantage of the setup is that the experiments were *person independent*, meaning that the same participant was never represented in both the training and test sets. This excludes the possibility of the approach recognizing the identity of the participants and not their condition.

Training the models required the setting of five hyperparameters, namely, *learning rate* (a factor that influences the size of the parameter updates during training), *number of neurons in the hidden layer*, *number of training epochs* (the number of cycles through which the network is trained), *batch size* (number of training samples used at any training epoch) and *padding* (length of the vector sequences fed to the networks). In this work, such a task was performed through *hyperparameter optimization*, meaning

**Table 3** The table shows, at both the clause (C) and person (P) level, accuracy, precision, recall and F1 measure. Since all experiments were repeated 30 times, the values are accompanied by their individual standard errors. JR and GMU stand for *joint representation* and *gated multimodal unit*, respectively

Modality	Level	Acc. (%)	Prec. (%)	Rec. (%)	F1 (%)
Text	C	60.4 ± 0.003	56.1 ± 0.005	46.5 ± 0.007	51.0 ± 0.005
Text	P	72.9 ± 0.020	100.0 ± 0.000	44.8 ± 0.040	61.9 ± 0.040
Audio	C	70.0 ± 0.006	65.1 ± 0.008	65.0 ± 0.008	65.0 ± 0.007
Audio	P	74.6 ± 0.021	76.9 ± 0.032	69.0 ± 0.306	72.7 ± 0.021
Sum Rule	C	66.4 ± 0.004	60.0 ± 0.006	54.3 ± 0.008	57.0 ± 0.005
Sum Rule	P	83.0 ± 3.1	95.2 ± 0.032	69.0 ± 0.067	80.0 ± 0.049
JR	C	64.0 ± 0.004	59.3 ± 0.006	55.2 ± 0.008	57.2 ± 0.005
JR	P	84.7 ± 0.027	95.4 ± 0.030	72.4 ± 0.043	82.3 ± 0.027
GMU	C	63.0 ± 0.004	58.1 ± 0.005	54.5 ± 0.010	56.2 ± 0.007
GMU	P	83.0 ± 0.024	95.2 ± 0.024	69.0 ± 0.049	80.0 ± 0.032

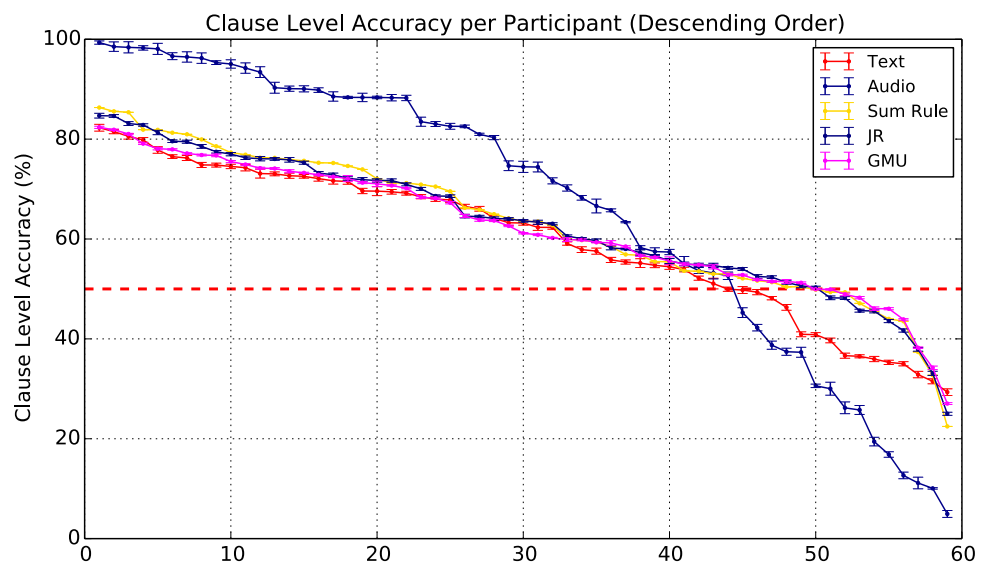
that the training set was split into two disjoint subsets with two-thirds and one-third of its samples, respectively (the smaller set is referred to as the *validation set*), and the automatic hyperparameter tuning approach with Bayesian optimization [46] was then applied to conduct a guided search for the best hyperparameters. The combination showing the highest accuracy over the validation set was retained and used to classify the samples of the test set.

During the experiments, the predefined sets used for the different hyperparameters were as follows: for the learning rate  $\lambda$ , the values were  $10^{-3}$ ,  $3 \times 10^{-3}$ ,  $10^{-2}$ , and  $10^{-1}$ . In the case of the hidden layers, the number of neurons was 32, 64 or 128. The training epochs were 30, 50 or 80, and the batch sizes were 32, 64 or 128 samples. Finally, the padding values for speech were 40, 50, 60, 70, 80, 100 and 120, while those for text were the integers between 9 and 14.

of data being used. The training was performed through backpropagation using the Adam optimizer and categorical cross-entropy as a loss function<sup>1</sup>.

Table 3 shows the depression detection results obtained with unimodal and multimodal approaches at the level of both clauses and participants. Given that the initialization of the network weights took place through a random process, every experiment was repeated 30 times, and the results are therefore reported in terms of average and standard deviation of the different performance metrics. The limited variance across the 30 repetitions suggests that the models are sufficiently robust to changes in the initialization, and the averages can therefore be considered realistic estimates of the performance. According to a two-tailed *t*-test with Bonferroni correction, the accuracy is always better than chance to a statistically significant extent at the level of both clauses and participants.

**Fig. 5** The figure shows, in descending order, the clause-level accuracy per participant. The curves corresponding to the multimodal approaches intersect the 50% horizontal line later. This means that correctly classified clauses tend to be distributed across a greater number of participants, and as a consequence, there is a greater number of cases in which the majority vote leads to a correct person classification. JR and GMU stand for *joint representation* and *gated multimodal unit*, respectively



In all cases, the main motivation behind the choice of the values was that they are considered standard in the literature. The only exception is padding, which depends on the type

<sup>1</sup> All models and training methodologies were implemented with TensorFlow.



**Table 4** The table shows the accuracy gain  $\Delta\alpha$  for the different approaches used in the experiments. The values  $\alpha_{min}$  and  $\alpha_{max}$  are the minimum and maximum accuracy that can result from the application of the majority vote, respectively

Modality	$\alpha_{min}$ (%)	$\alpha_{max}$ (%)	$\Delta\alpha$ (%)
Text	47.4	98.4	52.3
Audio	55.9	100	41.5
Sum Rule	49.1	99.5	63.5
JR	49.1	99.7	67.0
GMU	49.1	99.7	67.0

According to Table 3, multimodal approaches outperform unimodal ones in terms of person-level accuracy, the metric that matters from an application point of view. However, the speech-based unimodal approach shows the highest clause-level accuracy. Overall, this means that multimodal approaches benefit to a greater extent from the majority vote. Figure 5, showing the individual clause-level accuracies in descending order, provides a possible explanation of this observation. In particular, the figure shows that in the case of the multimodal approaches, correctly classified clauses tend to be more evenly distributed across participants. This leads to a greater number of cases in which the accuracy is above 50% (the condition for the majority vote to work).

While providing a possible explanation of why the majority vote is more beneficial for certain approaches, the observations above do not show to what extent the benefit can be considered satisfactory. One way to do so is to consider the accuracy gain  $\Delta\alpha$ :

$$\Delta\alpha = \frac{\alpha - \alpha_{min}}{\alpha_{max} - \alpha_{min}}, \tag{2}$$

where  $\alpha$  is the person-level accuracy actually observed after the majority vote, and  $\alpha_{min}$  and  $\alpha_{max}$  are the minimum and maximum person-level accuracy a majority vote can lead to. Person-level accuracy  $\alpha_{min}$  can be observed when all correctly classified clauses concentrate in the smallest possible number of participants. In contrast, the maximum value  $\alpha_{max}$  can be observed when the clause-level accuracy is the same for all participants. Given that the clause-level accuracy  $\alpha_c$  can be thought of as the probability of making the right decision about a clause,  $\alpha_{max}$  can be estimated as the probability of having more than half of the clauses classified correctly:

$$\alpha_{max} \approx \sum_{k=M/2+1}^M \binom{M}{k} \alpha_c^k (1 - \alpha_c)^{M-k}, \tag{3}$$

where  $M$  is the average number of clauses per participant (114 in the data of this work).

Table 4 shows the results of the different approaches; in particular, it shows that the multimodal ones tend to obtain

higher accuracy gains. In other words, achieving high clause-level accuracy is not sufficient to correctly classify the participants. It is also necessary for the distribution of correctly classified clauses to allow a clause accuracy higher than 50% for the largest possible number of participants. This is important because the networks are trained to maximize the clause-level accuracy but not to perform uniformly across participants. Therefore, there is a possible misalignment between the way the models are trained and the actual goal of the approaches. The next section shows how and when the combination of multiple modalities can address, at least to a partial extent, such a problem.

### Analysis of Multimodal Recognition

The previous section shows that the unimodal approaches tend to concentrate correctly classified clauses in a smaller number of participants. One possible explanation is that some of the participants tend to consistently manifest their condition through at least one of the modalities. In this way, they leave detectable traces of their condition in many clauses, and therefore, they make it easier to achieve high accuracy at the clause level. The participants who tend to do this through only one modality are likely to inject *diversity* [14], i.e., to lead the unimodal classifiers to make different mistakes over different participants. This can be an important advantage because a multimodal approach is beneficial mainly when unimodal approaches disagree; hence, one of these has a chance to compensate for the errors of the other.

Following up on the above, one possible way to measure the diversity is to compare  $N_d$ , the number of times the two unimodal approaches classify the same participant differently, with its upper bound, i.e., with the number  $N_{max}$  of disagreements expected when the two unimodal approaches are statistically independent. According to the data,  $N_d = 21$ , while  $N_{max}$  can be estimated as follows (the accuracy can be thought of as the probability of making the right decision about a participant):

$$N_{max} = [\alpha_1(1 - \alpha_2) + \alpha_2(1 - \alpha_1)]N, \tag{4}$$

where  $\alpha_1$  and  $\alpha_2$  are the person-level accuracies of the two unimodal approaches and  $N = 59$  is the total number of participants. Based on the results of Table 3,  $N_{max} = 23$ , meaning that  $N_d$  is 91.3% of its upper bound, and the unimodal approaches appear to be highly diverse.

These results suggest that a significant fraction of the participants ( $N_d$  corresponds to 35.6% of the total) tend to manifest their condition either through one modality or through the other. In particular, Table 5 shows that the depressed participants tend to manifest their pathology rather clearly through the way they speak while doing so

**Table 5** The table considers the 21 cases (out of the total of 59) for which there is disagreement between the two unimodal approaches. When the audio-based approach is the correct one, the classified participant is always depressed. In contrast, when the text-based approach is correct, the distribution of the participants across the classes is roughly uniform. One possible explanation is that when depressed people tend to manifest their condition through only one modality, they tend to do so through audio, i.e., through the way they speak

Correct Modality	Depressed	Control
Audio	11	0
Text	4	6

more ambiguously through the words they use (hence the high recall of the audio-based unimodal approach). As a confirmation, the GMU weights of the depressed subjects are higher for audio than for text (58.9% vs 54.9%), and the converse is true for the control subjects (45.0% vs 42.0%). The resulting diversity of the unimodal approaches is the probable reason why all multimodal systems show significantly higher person-level accuracy (see Table 3). In particular, the best multimodal approach performs better than the best unimodal system by 10.1 points. Furthermore, the person-level accuracy difference between multimodal and unimodal approaches is always statistically significant ( $p < 0.05$  according to a two-tailed  $t$ -test).

The problem left open in the above is to what extent the improvement resulting from the application of multimodal approaches can be considered satisfactory. For this reason, it is possible to estimate how close the performance of the multimodal approaches is to  $\alpha_{max}$ , the upper bound of the accuracy, which can be estimated as follows (it is the probability of at least one of the two unimodal approaches making the right decisions and, hence, giving the combination a chance to make the right decision too):

$$\alpha_{max} = 1 - (1 - \alpha_1)(1 - \alpha_2), \quad (5)$$

where  $\alpha_1$  and  $\alpha_2$  are the person-level accuracies of the unimodal approaches. The value of  $\alpha_{max}$  is 93.2% and, therefore, the person-level accuracy of the multimodal approaches ranges between 89.0% and 90.9% of  $\alpha_{max}$ , the maximum that can be obtained with the two unimodal approaches at disposition. In particular, given that the best unimodal approach achieves an accuracy of 74.6%, the improvement by 10.1 points (see above) corresponds to 54.3% of the maximum improvement that can be achieved. In other words, roughly half of the times when there is disagreement between the two modalities, the one leading to the correct classification compensates for the error of the other.

## Confidence Measures

The results presented so far suggest that the proposed approaches can make the right decision about an individual approximately 4 times out of 5, but it is unclear whether this proportion can be considered satisfactory. One possible benchmark for comparison is the performance of *general practitioners* (GPs), the doctors who are the first line of intervention against depression, especially in regard to convincing possible patients to seek treatment. According to a meta-analysis of the literature, the *sensitivity*<sup>2</sup> and *specificity*<sup>3</sup> of GPs are in the ranges of 41.3% to 59.0% and 74.5% to 87.3%, respectively [47]. This corresponds to an accuracy between 57.9% and 73.1% for the data used in the experiments of this work.

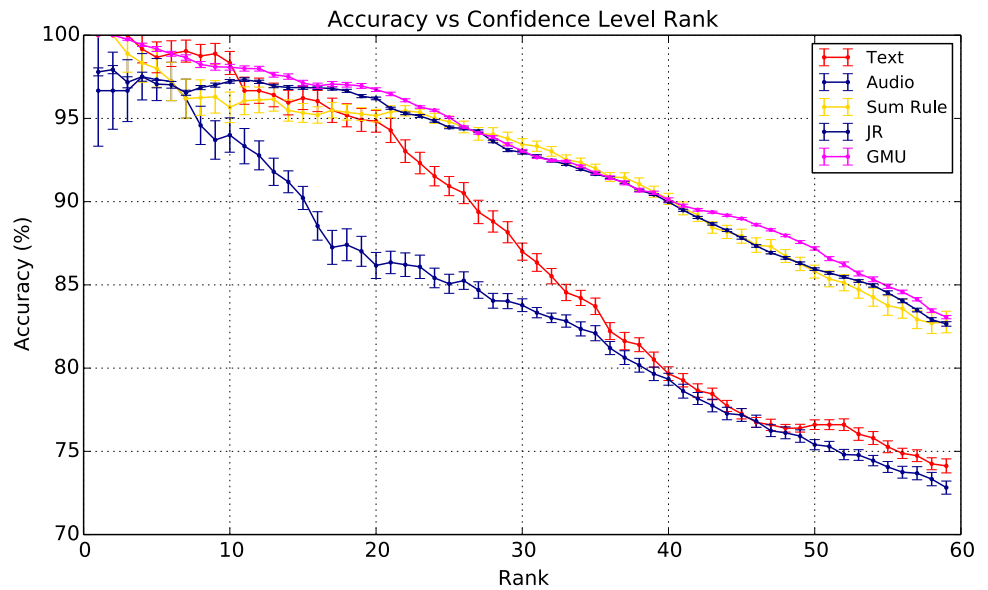
According to the above, all approaches proposed in this work appear to perform comparably to an average GP, especially in terms of sensitivity (the name of recall in medical domains). Such a measure is particularly important because Type I errors (classifying a depressed person as nondepressed) are those that have the most negative consequences and, therefore, should be as limited as possible. This suggests that one possible approach to the application of depression detection technologies is to identify cases in which the outcome of a system can be trusted while leaving the others to medical attention. This appears to be in line with recent trends suggesting that AI-driven technologies should collaborate with users and not simply replace them [48].

One possible way to address the problem above is to consider only those participants for whom the two unimodal approaches agree with each other. The rationale is that agreement between multiple modalities might correspond to higher confidence and, correspondingly, to higher performance. In the experiments of this work, the unimodal approaches agree 38 times out of 59 (corresponding to 64.4% of the total participants): 24 times for a control participant and 14 for a depressed one. In 33 of the 38 cases, both approaches are correct (corresponding to an accuracy of 86.8%). In the remaining 5 cases, the participants are always depressed, thus resulting in a recall of 64.3%. This means that filtering the participants according to agreement between modalities increases accuracy while keeping the sensitivity at the level of an average GP. As a consequence, at least in the experiments of this work, it is possible to automatically process roughly two-thirds of the participants while leaving only the remaining third to the doctors (without accuracy or sensitivity losses compared to the doctors considering all participants).

<sup>2</sup> Percentage of depressed individuals actually diagnosed as such (equivalent to recall).

<sup>3</sup> Percentage of nondepressed individuals actually diagnosed as such.

**Fig. 6** The plots show the accuracy when taking into account only the  $r$  persons with the highest confidence values. On average, multimodal approaches appear to have higher accuracy for every value of  $r$ , and in particular, they appear to have at least 90% accuracy when taking into account the 40 top-ranking participants. In other words, it is possible to automatically isolate two-thirds of the participants for whom the system decides correctly 9 times out of 10. JR and GMU stand for *joint representation* and *gated multimodal unit*, respectively



The main disadvantage of the approach above is that it can be applied only to the unimodal approaches that, according to Table 3, have the lowest performance. For this reason, it is necessary to define a confidence measure that is independent of the particular approach being used. One possibility is to consider the following:

$$c = \frac{\max(n_D, N - n_D)}{N} \tag{6}$$

where  $N$  is the total number of clauses a participant has uttered and  $n_D$  is the number of clauses that, for a given participant, have been assigned to the *depression* class. The *rationale* behind the definition above is that the higher the fraction of clauses the approach assigns to a given class, the higher the confidence of the system.

The measure above allows one to rank the participants according to the value of  $c$  (from largest to smallest) and to consider the accuracy at position  $r$ . If higher values of  $c$  actually correspond to correct decisions, the accuracy should be high when taking into account only the top positions of the ranking. Figure 6 appears to confirm such an expectation, and in particular, it shows that the multimodal approaches have an accuracy higher than 90% when taking into account the 40 top-ranking participants (roughly two-thirds of the total). In this respect, the approach appears to be able to discriminate between cases that are sufficiently clear to be processed automatically and cases that require medical attention, thus allowing the system to potentially reduce the workload of the medical personnel by two-thirds while still keeping the accuracy above 90%.

### Discussion and Conclusions

This article has presented depression detection experiments performed over a corpus of interviews involving 29 depression patients and 30 persons who had never experienced mental health issues. The main limitation of the work is that the interview transcriptions were obtained manually, but this corresponds to a common practice in state-of-the-art depression detection ([Survey of Previous Work](#) shows that some of the most important benchmarks in the literature make use of manual transcriptions). However, the distinction between the depressed and control participants was made by professional psychiatrists and not through the administration of self-assessment questionnaires. This is an important advantage because it makes it more likely that the proposed approaches actually learned to detect depression. Furthermore, [The Data](#) shows that, at least in the case of this work, the BDI-II scores are not fully reliable, especially in regard to depression patients.

Overall, the results show that it is possible to achieve an accuracy of more than 80%, roughly corresponding to 4 correct decisions out of 5. According to the literature [47], such a result corresponds to the average performance of a general practitioner, typically the first doctor to spot depression and possibly to stimulate potential patients to contact a psychiatrist. In this respect, the experiments illustrate several application scenarios where the proposed approaches use confidence measures that identify the cases most likely to be correctly classified. In this way, the systems

can reduce the workload of doctors by up to two-thirds while still ensuring the desired level of performance. This is important because it can increase the efficiency of screening services and, correspondingly, reduce the costs associated with depression diagnosis.

The analysis of the performance shows that one of the main differences between unimodal and multimodal methodologies is that the latter tend to have more uniform clause-level accuracy across the participants (see [Experiments and Results](#)). This is important because it leads to higher person-level accuracy, the metric that actually matters from an application point of view (see [Fig. 5](#)). Such a result stems from the tendency of certain participants, in particular depressed ones, to manifest their condition either through what they say or through how they say it but not through both. To the best of our knowledge, such an observation has not previously been made in the literature, but it is an important aspect of this work. The reason is that it is a source of diversity across the unimodal approaches, and it is owing to such a property that these approaches disagree about a participant roughly one-third of the times. In this way, the correct unimodal approach has a chance to compensate for the error of the other, the key assumption underlying multimodal methodologies.

The observations above suggest that it is the behavior of the participants, at least to a certain extent, that determines the conditions under which the approaches will work. This is important because it might explain why state-of-the-art methods are uncertain in identifying the best way to detect depression (see [Survey of Previous Work](#)). In fact, the way people manifest depression can change significantly from one individual to another depending on a large number of social, psychological, economic and cultural factors [1]. As a result, none of the behaviors considered in the literature (facial expressions, paralinguistics, body movements, etc.) appear to clearly outperform the others. The main reason why this work focuses on linguistic and acoustic aspects of speech is that depression interferes with the neural processes underlying language and communication (see, e.g., [49, 50]), thus leaving detectable traces in both what people say and how they say it. In addition, the use of speech has several advantages from an application point of view, including the possibility of detecting depression via phone [25], typically the means through which people contact counseling services, or using ordinary laptop microphones in an informal setting, as was the case for the data used in this work.

According to the Gartner Group, one of the most important strategic consulting companies in the world, the detection of mental health issues is one of the most promising areas of Social and Emotion AI ([www.gartner.com/smarterwithgartner/13-surprising-uses-for-emotion-ai-technology/](http://www.gartner.com/smarterwithgartner/13-surprising-uses-for-emotion-ai-technology/)), the AI areas concerned with the inference of affective phenomena from observable data. The main

reason is the increasingly greater number of people affected by mental health issues [3] and the resulting pressure on healthcare services. In such a context, approaches such as those presented in this work can support the work of psychiatric and counseling services, possibly allowing doctors to concentrate on ambiguous and difficult cases while leaving machines to deal with the most evident ones. For these reasons, besides improving current methodologies and collecting more data, future work will focus on the development of strategies aimed at integrating technology into clinical practice, in line with observations showing that the best way to apply AI is to use it to support humans and not to replace them [48].

In addition to the above, future work will target the integration of the proposed approach into clinical practice. In this respect, a possible protocol includes the following steps:

- Step 1: Recording of the interviews;
- Step 2: Automatic analysis of the interview recordings;
- Step 3: Discussion of the analysis outcome among psychiatrists;
- Step 4: Acceptance or rejection of the automatic analysis outcome.

For Step 1, the main recommendation is to define a stable setting, meaning that the interviews should always take place in the same place and always with the same equipment. Furthermore, the position of the participants with respect to the computer and microphone used to record the interviews should not change from one person to another. For step 2, the suggestion is to implement the algorithms with publicly available libraries such as TensorFlow<sup>4</sup> so that the results of the analysis will always be reproducible by third parties. In a similar vein, one further recommendation is to make the code publicly available so that potential flaws or errors can be identified by a potential community of users. Step 3 corresponds to normal practice in psychiatric hospitals, where all cases are discussed collegially among doctors of the same department. Step 4 is important because it leaves to the doctors the responsibility of accepting or rejecting the outcome of the automatic analysis. This ensures that AI acts as a tool aimed to improve efficiency and effectiveness and not as a technology that replaces doctors. One possible recommendation for this step is to provide medical personnel with training in AI literacy, i.e., in developing a level of technical competence sufficient to integrate AI into professional practice. In particular, the training should stimulate awareness that the outcome of automatic approaches is subject to errors and should be considered competent advice rather than a fully accurate judgment.

<sup>4</sup> <https://www.tensorflow.org>

**Funding** The research leading to these results received funding from the project ANDROIDS funded by the program V:ALERE 2019 Università della Campania Luigi Vanvitelli, D.R. 906 del 4/10/2019, prot. n. 157264,17/10/2019. The work of Alessandro Vinciarelli was supported by UKRI and EPSRC through grants EP/S02266X/1 and EP/N035305/1, respectively.

## Declarations

**Conflicts of Interest** Authors Nujud Alosbhan, Anna Esposito and Alessandro Vinciarelli declare that they have no conflict of interest.

**Ethical Approval** All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. Informed consent was obtained from all individual participants included in the study. Participants accepted being involved in the experiments on a voluntary basis and signed an informed consent agreement formulated in accord with the privacy and data protection procedures established by Italian and European laws. The ethical committee of the Department of Psychology at Università degli Studi della Campania, “Luigi Vanvitelli”, authorized this research with protocol number 09/2016.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

- Irons C. Depression. Palgrave (2014).
- Marcus M, Yasamy T, Ommeren M, Chisholm D, Saxena S. Depression: A global health concern. World Federation for Mental Health: Tech. rep; 2012.
- WHO Document Production Services. Depression and other common mental disorders. Tech. rep.: World Health Organization; 2017.
- Wang P, Angermeyer M, Borges G, Bruffaerts R, Chiu W, De Girolamo G, Fayyad J, Gureje O, Haro J, Huang Y, Kessler R, Kovess V, Levinson D, Yoshitomi N, Oakley Brown M, Ormel J, Posada-Villa J, Aguilar-Gaxiola S, Alonso J, Lee S, Heeringa S, Pennell B, Chatterji S, Bedirhan Üstün T. Delay and failure in treatment seeking after first onset of mental disorders in the world health organization’s world mental health survey initiative. *World Psychiatry*. 2007;6(3):177–85.
- Vinciarelli A, Pantic M, Bourlard H. Social Signal Processing: Survey of an emerging domain. *Image Vis Comput*. 2009;27(12):1743–59.
- Picard RW. Towards computers that recognize and respond to user emotion. *IBM Systems Journal*. 2000;39(3–4):705–19.
- Schuller B, Batliner A. Computational paralinguistics: emotion, affect and personality in speech and language processing. John Wiley & Sons; 2013.
- Baltrušaitis T, Ahuja C, Morency LP. Multimodal machine learning: A survey and taxonomy. *IEEE Trans Pattern Anal Mach Intell*. 2018;41(2):423–43.
- Arevalo J, Solorio T, Montes-y Gómez M, González F. Gated Multimodal Units for information fusion. arxiv:1702.01992.
- Rohanian M, Hough J, Purver M. Detecting depression with word-level multimodal fusion. In: Proceedings of Interspeech; 2019. p. 1443–7.
- Williamson J, Godoy E, Cha M, Schwarzentruher A, Khorrami P, Gwon Y, Kung HT, Dagli C, Quatieri T. Detecting depression using vocal, facial and semantic communication cues. In: Proceedings of the International Workshop on Audio/Visual Emotion Challenge; 2016. p. 11–8.
- Pentland A. *Honest Signals*. MIT Press; 2007.
- Paulhus D, Vazire S. The self-report method. In: R. Robins, R. Fraley, R. Krueger (eds.) *Handbook of Research Methods in Personality Psychology*. Gilford; 2007. p. 224–39.
- Ranawana R, Palade V. Multi-classifier systems: Review and a roadmap for developers. *International Journal of Hybrid Intelligent Systems*. 2006;3(1):35–61.
- Valstar M, Schuller B, Krajewski J, Cohn J, Cowie R, Pantic M. AVEC 2014 – The three dimensional affect and depression challenge. In: Proceedings of the ACM International Workshop on Audio/Visual Emotion Challenge; 2014. p. 1–9.
- Valstar M, Schuller B, Smith K, Eyben F, Jiang B, Bilakhia S, Schnieder S, Cowie R, Pantic M. AVEC 2013: The continuous audio/visual emotion and depression recognition challenge. In: Proceedings of the ACM International Workshop on Audio/visual Emotion Challenge; 2013. p. 3–10.
- Ringeval F, Schuller B, Valstar M, Gratch J, Cowie R, Scherer S, Mozgai S, Cummins N, Schmitt M, Pantic M. AVEC 2017: Real-life depression, and affect recognition workshop and challenge. In: Proceedings of the International Workshop on Audio/Visual Emotion Challenge; 2017. p. 3–9.
- Valstar M, Gratch J, Schuller B, Ringeval F, Lalanne D, Torres Torres M, Scherer S, Stratou G, Cowie R, Pantic M. AVEC 2016: Depression, mood, and emotion recognition workshop and challenge. In: Proceedings of the International Workshop on Audio/Visual Emotion Challenge; 2016. p. 3–10.
- Gilbody S, Richards D, Brealey S, Hewitt C. Screening for depression in medical settings with the Patient Health Questionnaire (PHQ): A diagnostic meta-analysis. *J Gen Intern Med*. 2007;22(11):1596–602.
- Alghowinem S, Goecke R, Wagner M, Parker G, Breakspear M. Head pose and movement analysis as an indicator of depression. In: Proceedings of the IEEE International Conference on Affective Computing and Intelligent Interaction; 2013. p. 283–8.
- Cai H, Zhang X, Zhang Y, Wang Z, Hu B. A case-based reasoning model for depression based on three-electrode eeg data. *IEEE Trans Affect Comput (to appear)*; 2019.
- Girard J, Cohn J, Mahoor M, Mavadati S, Rosenwald D. Social risk and depression: Evidence from manual and automatic facial expression analysis. In: Proceedings of the IEEE International Conference and Workshops on Automatic Face and Gesture Recognition; 2013. p. 1–8.
- Joshi J, Goecke R, Parker G, Breakspear M. Can body expressions contribute to automatic depression analysis? In: Proceedings of the IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG); 2013. p. 1–7.
- Yang Y, Fairbairn C, Cohn J. Detecting depression severity from vocal prosody. *IEEE Trans Affect Comput*. 2012;4(2):142–50.
- Huang Z, Epps J, Joachim D, Chen M. Depression detection from short utterances via diverse smartphones in natural environmental conditions. In: Proceedings of Interspeech; 2018. p. 3393–7.
- Low LA, Maddage NC, Lech M, Sheeber L, Allen N. Detection of clinical depression in adolescents’ speech during family interactions. *IEEE Trans Biomed Eng*. 2011;58(3):574–86.

27. Cummins N, Sethu V, Epps J, Schnieder S, Krajewski J. Analysis of acoustic space variability in speech affected by depression. *Speech Communication*. 2015;75:27–49.
28. Morales M, Levitan R. Speech vs. text: A comparative analysis of features for depression detection systems. In: proceedings of the IEEE Spoken Language Technology Workshop; 2016. p. 136–43.
29. Alhanai T, Ghassemi M, Glass J. Detecting depression with audio/text sequence modeling of interviews. In: Proceedings of Interspeech; 2018.
30. Andrade L, Caraveo-Anduaga J, Berglund P, Bijl R, De Graaf R, Vollebergh W, Dragomirecka E, Kohn R, Keller M, Kessler R, Kawakami N, Kiliç C, Offord D, Bedirhan Ustun T, Wittchen HU. The epidemiology of major depressive episodes: Results from the international consortium of psychiatric epidemiology (ICPE) surveys. *Int J Methods Psychiatr Res*. 2003;12(1):3–21.
31. Kessler R, Berglund P, Demler O, Jin R, Koretz D, Merikangas K, Rush A, Walters E, Wang P. The epidemiology of major depressive disorder: Results from the National Comorbidity Survey Replication (NCS-R). *J Am Med Assoc*. 2003;289(23):3095–105.
32. Garber J, Gallerani C, Frankel SA. Depression in children. In: I. Gotlib, C. Hammen (eds.) *Depression in children*. The Guilford Press; 2009. p. 405–43.
33. Kessler R, Walters E. Epidemiology of DSM-III-R major depression and minor depression among adolescents and young adults in the national comorbidity survey. *Depress Anxiety*. 1998;7(1):3–14.
34. McDougall F, Matthews F, Kvaal K, Dewey M, Brayne C. Prevalence and symptomatology of depression in older people living in institutions in England and Wales. *Age Ageing*. 2007;36(5):562–8.
35. Bos E, Bouhuys A, Geerts E, Van Os T, Ormel J. Lack of association between conversation partners' nonverbal behavior predicts recurrence of depression, independently of personality. *Psychiatry Res*. 2006;142(1):79–88.
36. Geerts E, Bouhuys N, Van den Hoofdakker R. Nonverbal attunement between depressed patients and an interviewer predicts subsequent improvement. *J Affect Disord*. 1996;40(1–2):15–21.
37. Beck A, Alford B. *Depression: Causes and Treatment*. University of Pennsylvania Press; 2009.
38. Graves A. *Supervised Sequence Labelling with Recurrent Neural Networks*. Springer Verlag; 2012.
39. Kittler J, Hatef M, Duin R, Matas J. On combining classifiers. *IEEE Trans Pattern Anal Mach Intell*. 1998;20(3):226–39.
40. El Ayadi M, Kamel M, Karray F. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recog*. 2011;44(3):572–87.
41. Vinciarelli A, Mohammadi G. A survey of personality computing. *IEEE Trans Affect Comput*. 2014;5(3):273–91.
42. Yamada I, Asai A, Shindo H, Takeda H, Takefuji Y. Wikipedia2Vec: An optimized implementation for learning embeddings from Wikipedia. arXiv preprint 2018. arXiv:1812.06280.
43. Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. Distributed representations of words and phrases and their compositionality. In: *Adv Neural Inf Proces Syst*; 2013. p. 3111–9.
44. Charniak E. *Introduction to Deep Learning*. MIT Press; 2018.
45. Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint 2018. arXiv:1810.04805.
46. Osborne M. *Bayesian Gaussian processes for sequential prediction, optimisation and quadrature*. Ph.D. thesis, Oxford University, UK. 2010.
47. Mitchell A, Vaze A, Rao S. Clinical diagnosis of depression in primary care: a meta-analysis. *The Lancet*. 2009;374(9690):609–19.
48. Davenport T. *The AI advantage: How to put the artificial intelligence revolution to work*. MIT Press; 2018.
49. D'Arcy R, Connolly J, Service E, Hawco C, Houlihan M. Separating phonological and semantic processing in auditory sentence processing: A high-resolution event-related brain potential study. *Hum Brain Mapp*. 2004;22(1):40–51.
50. Seghier M, Lazeyras F, Pegna A, Annoni JM, Zimine I, Mayer E, Michel C, Khateb A. Variability of fMRI activation during a phonological and semantic language task in healthy subjects. *Hum Brain Mapp*. 2004;23(3):140–55.