Demiroglu *et al. EURASIP Journal on Audio, Speech, and Music Processing* (2020) 2020:17
https://doi.org/10.1186/s13636-020-00182-4

EURASIP Journal on Audio,
Speech, and Music Processing

RESEARCH Open Access

Depression-level assessment from multi-lingual conversational speech data using acoustic and text features

Cenk Demiroglu[1*], Aslı Beşirli[2], Yasin Ozkanca[1] and Selime Çelik[2]

Abstract

Depression is a widespread mental health problem around the world with a significant burden on economies. Its early diagnosis and treatment are critical to reduce the costs and even save lives. One key aspect to achieve that goal is to use technology and monitor depression remotely and relatively inexpensively using automated agents. There has been numerous efforts to automatically assess depression levels using audiovisual features as well as text-analysis of conversational speech transcriptions. However, difficulty in data collection and the limited amounts of data available for research present challenges that are hampering the success of the algorithms. One of the two novel contributions in this paper is to exploit databases from multiple languages for acoustic feature selection. Since a large number of features can be extracted from speech, given the small amounts of training data available, effective data selection is critical for success. Our proposed multi-lingual method was effective at selecting better features than the baseline algorithms, which significantly improved the depression assessment accuracy. The second contribution of the paper is to extract text-based features for depression assessment and use a novel algorithm to fuse the text- and speech-based classifiers which further boosted the performance.

Keywords: Depression detection, Acoustic features, Feature selection

1 Introduction

Depression is a vital problem that affects a large percentage of the population around the world. It not only affects the well-being and productivity of individuals but also causes heavy economic burden on the society [1]. In fact, with more than 300 million depression patients, world health organization (WHO) declared depression as the leading cause of ill health and disability worldwide [2]. Because access to the diagnosis and treatment are expensive and sometimes not possible, inexpensive and accurate diagnosis with the help of technology became an increasingly important research challenge [3].

Speech signal has been investigated for detecting depression since it carries significant information about mental health of the speakers [4–7]. Combined with the pervasive use of smartphones in our daily lives, hence, relatively easy and non-intrusive access to good-quality speech data, remote monitoring of patients through acoustic analysis became a promising research area [8].

In [9], phase distortion deviation that is used for voice quality examinations is found to be helpful for detecting depression. In [10], distortions in formant trajectories were used to detect depression. In [11], degradation in spectral variability was used. In [12], gender-dependent feature extraction was found to improve the detection performance. In [13], i-vectors and MFCC features that are commonly used for speaker verification were found to be helpful for depression detection even when the utterances were only 10 s long.

In AVEC 2019 [14], depression detection subchallenge, mel-frequency cepstral coefficients (MFCC) features, and extended GeMAPS [15] features were extracted from

*Correspondence: cenk.demiroglu@ozyegin.edu.tr
†Equal contributors
[1] Department of Electrical and Electronics Engineering, Ozyegin University, Istanbul, Turkey
Full list of author information is available at the end of the article

© The Author(s). 2020 **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/.

audio. Those features were combined within a bag-of-word (BoW) front-end that uses vector quantization to quantize feature vectors into a limited set of centroids [16]. Thus, sequences of features vectors are converted into sequences of centroid-ids. The method is well-known in the text processing field, and it has also gained recent popularity in emotion detection [17].

Deep learning based feature extraction methods for audio analysis also gained recent popularity in emotion detection [14]. In that approach, convolutional neural network (CNN)-based image recognition systems are provided with speech spectrograms and resulting activation values are used for emotion detection [18].

Besides the speech signal, there are also audio-visual methods for detecting depression. In [19], face analysis and speech prosody are used for depression detection. Similarly, audio-visual features are used in [20–23]. Retardations in motor control due to depression causes changes in coordination and timing of speech and face movements, which are used for audio-visual detection in [24].

This paper has two contributions. One of the contributions is novel algorithms for multi-lingual feature selection where three databases, Turkish, German, and English were used together to improve redundancy and relevance computations in the case of data sparsity. The second contribution is a novel feature fusion technique where transcription-based predictions were used to adjust the predictions of the acoustic-only model when the predictions of those two modalities were highly conflicting. Significant improvements are obtained for the Turkish, German, and English databases using the proposed techniques.

2 Related work

2.1 Feature selection

A large number of acoustic features can be derived from conversational speech to detect depression. However, building models with those features is challenging because of the curse of dimensionality and the typically small amounts of training data available in depression studies.

One way of reducing the dimensionality of features is to use feature selection where features that are most relevant for the classification task and least correlated among themselves are selected for classification. To that end, Minimum Redundancy Maximum Relevance (MRMR) algorithm is commonly used [25–27].

In [28], a two-step feature selection algorithm was proposed. The conversation is segmented into topics and features are extracted for each topic. As a first step, correlation-based feature subset selection was applied regardless of the topics [29]. In the second step, the selected features for each topic were further refined by first ranking them based on relevance and selecting subsets using regression tests. In [30], a simple t test was used to select features from a set of 504 acoustic features.

Besides selecting features automatically, there are knowledge-based set of features that are designed for emotion detection. One of the more popular examples to that approach is the Geneva feature set (GeMAPS) [31] which is developed by augmenting a minimum set of acoustic features that were shown in the literature to be reliable indicators of emotional state and that have the highest theoretical significance.

2.2 Fusion of text and audio features

Transcriptions of the speech signal have also been used as another mode of information [3] for depression detection. In [32], transcription-derived features were used in addition to the speech features. Furthermore, sentiment analysis was performed on text and sentiment features were used to build an independent detector. Then, score fusion was used to combine acoustic and text-based system scores. Syntactic and semantic features were derived from transcriptions in [33] and shown to be effective indicators of depression.

Conversations with patients can be designed in a way to obtain data that is more indicative of depression, as opposed to a regular conversation. In [34], type of questions (positive and negative stimulus) during conversations have been shown to impact voice quality parameters in psychologically distressed subjects. Speech segments with higher articulation effort were found to be more informative for depression detection in [6].

In [35], biomarkers that are derived from facial coordination and timing features were used together with vocal cues and semantic features from dialogue content using a sparse-coded lexical embedding space. In [36], depressed individuals were shown to use less social words and more anxiety-related words.

A depression-detection algorithm is presented in [37] where interactions between subjects and the computer agent were modeled without explicit topic modeling. Long-short term memory (LSTM) neural networks were used with audio and text features. The results in [37] suggested that minimal knowledge of the conversation is required for depression detection.

In [38], both conversation-level (number of sentences, number of words used, etc.) and content-level (feeling good/bad, extrovert/introvert personality, etc.) information derived from the transcripts of the dialogs were used to extract features and then scores from both audio and text features were fused via a DNN model.

There are also attempts to extract both audio and text features using deep networks as well as fusing those features using a deep network. For example, deep spectrum features [14] for audio was fused with BERT-based text representation in [39] using fully connected layer.

2.3 Cross-lingual depression detection

In depression detection, a less studied research challenge is to use speech data from other languages to train models. This approach is not only important for understanding universal cues of depression across different cultures/languages, but it also allows the use of data from other languages, which is important given the typically small amounts of data available in the public databases for each language. In [40], prediction models built with a German database were shown to produce prediction scores in English that were correlated with the self-assessment scores. In [30], combination of datasets in different languages was shown to yield high accuracy whereas if the train and test data are in different languages, performance was found to be lower.

In [41], transfer of models developed for the resource-rich English language to other languages with limited datasets was investigated. The method was shown to be improve Aphasia detection and have promise for Alzheimer's disease detection.

3 Minimum redundancy maximum relevance (MRMR) feature selection

In the MRMR approach, F-statistic is used for computing the relevance of a selected feature set (S) for a K-class classification task. F-statistic for feature g_i is defined as

$$F(g_i) = \left[\frac{1}{\sigma_i^2 (K-1)} \sum_{k=1}^{K} n_k \left(\bar{g}_{i,k} - \bar{g}_i \right)^2 \right] \tag{1}$$

where $\bar{g}_{i,k}$ is the mean of g_i for the training samples in class k, $\bar{g}_i$ is the global mean of g_i over all samples in all classes, and n_k is the total number of samples in class-k. σ_i^2, the pooled variance, is

$$\left[\frac{1}{N_s - K} \sum_{k=1}^{K} (n_k - 1)\, \sigma_{i,k}^2 \right], \tag{2}$$

where $\sigma_{i,k}^2$ is the variance of g_i in class-k, and N_s is the total number of samples .

Relevance of a feature set S is then defined as

$$V_F(S) = \frac{1}{|S|} \sum_{i \in S} F(g_i). \tag{3}$$

Redundancy of the feature set S is defined using the Pearson's correlation for every possible feature combination:

$$W_c(S) = \frac{1}{|S|^2} \sum_{i,j} |c(i,j)|, \tag{4}$$

where $|c(i,j)|$ is the absolute value of the correlation $c(i,j)$ between feature i and feature j. Finally, the MRMR algorithm selects the features set (S) using

$$\arg\max_{S} \left\{ V_F(S) - W_c(S) \right\}. \tag{5}$$

4 Proposed feature selection algorithms

The MRMR algorithm works well in many machine learning problems. However, for the depression detection problem, training data is typically limited, and therefore, computation of the F-statistic and feature correlations are often unreliable. Here, we propose three algorithms to more reliably compute the statistics required for the MRMR algorithm as described below.

4.1 Multi-lingual computation of relevance

The F-statistic computation in Eq. (1) requires estimation of the global variance (σ_i^2), the global mean (g_i), and the class means ($\bar{g}_{i,k}$) for each class k and feature i. Even though the global mean and variance can usually be estimated relatively reliably, estimating the class means is more challenging when the number of classes is large and the data is limited as is often the case in depression screening tests.

The publicly available databases used in depression studies typically have less than 200 subjects. Moreover, commonly used depression evaluation tests BDI-II and PHQ-8 have 64 and 25 classes, respectively. Thus, the number of subjects available per class is usually not enough to compute the relevance reliably. In the multi-lingual MRMR (ml-MRMR) approach, to increase the number of available samples for each class and improve the computation of F-statistic, we propose populating each class using samples available in a different language for that same class. For example, if there is only one subject with a PHQ-8 score of 10 in the Turkish dataset, then feature vectors of subjects with a PHQ-8 score of 10 from the German datasets can be used to populate the Turkish dataset.

In some cases, the number of samples is still low after cross-lingual population of classes. In that case, samples from the neighboring classes in a different language are used for further increase the sample size. This approach takes advantage of the fact that subjects in neighboring classes (PHQ-8 levels 9 and 10, for example) are expected to be similar to each other. That assumption, though, becomes less valid, as the neighboring class is further away from the target class. Thus, while populating classes with cross-lingual data, each sample borrowed from the neighboring class is weighted according to its distance from the target class. The weight parameter γ is defined as

$$\gamma_j = e^{-j^2}. \tag{6}$$

where

$$j = |c_{tar} - c_{nb}| \tag{7}$$

is the distance of the target class c_{tar} to the neighboring class c_{nb}.

After cross-lingual population of class k, number of samples, n_k, is computed using the weight parameter γ as follows:

$$n_k^{'} = n_k + \sum_{j=-J_k}^{+J_k} \gamma_j \Gamma_{k+j} \tag{8}$$

where Γ_{k+j} is the number of samples borrowed from class $k+j$. J_k is set such that $n_k^{'} > N_{min}$. Thus, by including data from the same and neighboring classes in a different database, we ensure that there are at least N_{min} samples for each class in the target database. The adjusted mean of each class k, $\bar{g}_{i,k}^{'}$, is then

$$\bar{g}_{i,k}^{'} = \frac{1}{n_k^{'}} \left[n_k \bar{g}_{i,k} + \Sigma_{i,k} \right] \tag{9}$$

where the cross-lingual component

$$\Sigma_{i,k} = \sum_{j=-J_k}^{+J_k} \sum_{s=0}^{n_{k-j}} \gamma_j g_{i,k,j}(s) \tag{10}$$

and $g_{i,k,j}(s)$ is the ith feature of sample s borrowed from the jth neighbor of class k.

Using the $n_k^{'}$, and $\bar{g}_{i,k}^{'}$ the new F-score is

$$F^{'}\left(g_i\right) = \left[\frac{1}{\sigma_i^2 (K-1)} \sum_{k=1}^{K} n_k^{'} \left(\bar{g}_{i,k}^{'} - \bar{g}_i \right)^2 \right]. \tag{11}$$

An example of how sparse classes are populated is shown in Fig. 1 where the Turkish dataset is populated with samples from the German dataset. Note that the classes that do not have samples are not populated in the ml-MRMR algorithm as shown in Fig. 1. Those classes are ignored in the MRMR computations.

Figure 2 shows the histograms of $\bar{g}_{i,k}$ for all features i and classes k using the baseline MRMR and the proposed ml-MRMR algorithms with $N_{min} = 3$. Samples from the German database are used to populate the Turkish database. The distribution gets closer to a Gaussian with the ml-MRMR algorithm compared to the baseline MRMR algorithm. Moreover, heavy-tails generated with the baseline MRMR algorithm are suppressed, which indicates that the ml-MRMR algorithm can effectively reduce the outliers in the data.

4.2 Clustering approach

Depression screening tests often have large number of classes. For example, for PHQ-8 has 24 classes and BDI-II has 64 classes. However, in diagnosis, level of depression (severe, moderate, etc.) corresponds to a range of classes. For example, all subjects that have PHQ-8 scores between 20 and 24 are diagnosed as severely depressed subjects. Thus, the distinction between classes with similar scores is likely not represented in conversational speech. For instance, the difference between two subjects with scores of s or $s+1$ may not be as significant to warrant different classes for those two cases. Given the limited training data available, we propose clustering samples that have similar scores and reducing the number of distinct classes to increase the number of samples per class.

In the clustering approach, the depression classes are clustered the number of classes in the MRMR training process is reduced to improve the feature selection performance by increasing the data available for each class. In this approach, data is split uniformly into N_{clus} classes. Cluster centroids are found by first uniformly dividing the score scale. If the centroid class has no samples, then the

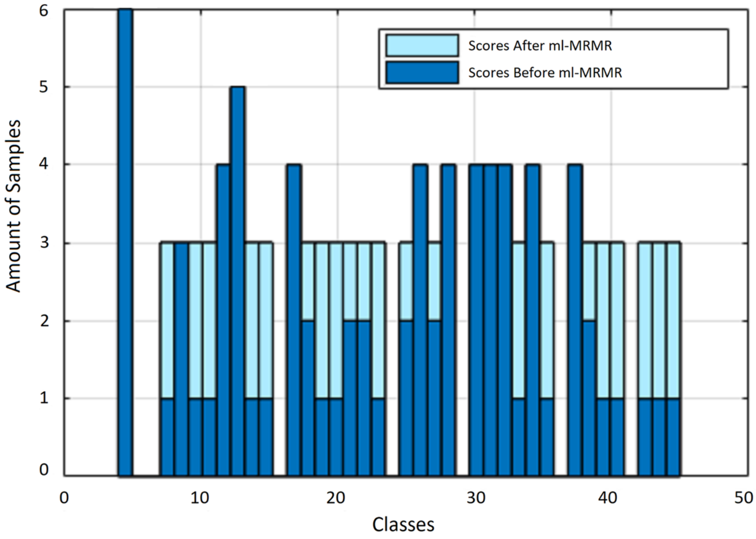

Fig. 1 Sample distribution before and after applying the ml-MRMR algorithm with $N_{min} = 3$. Turkish database is used

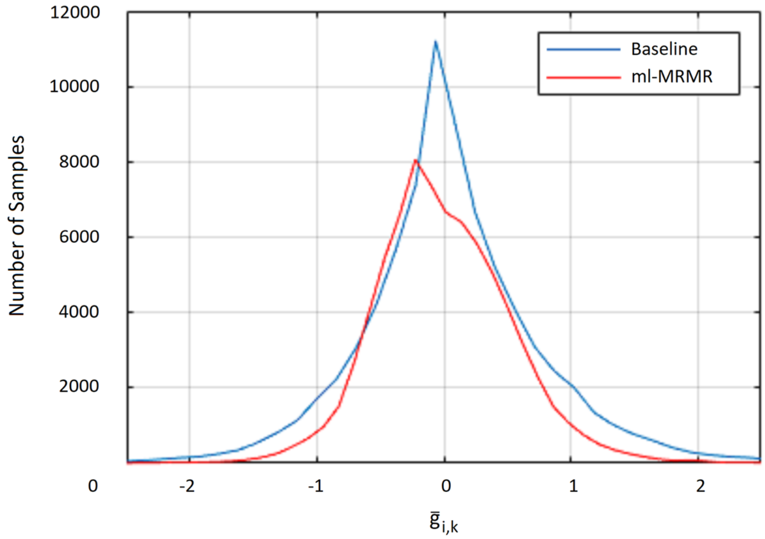

Fig. 2 Comparison of the distributions of $\bar{g}_{i,k}$ for the baseline MRMR and the proposed ml-MRMR algorithms. Turkish database is used

nearest non-empty class is assigned as the centroid. After setting the centroids, each class is assigned to the nearest centroid.

Figure 3 shows the sample distribution after the clustering approach is applied to the Turkish database with $N_{clus} = 14$. Comparing the new distribution to the original distribution in Fig. 1, distribution of samples per class becomes more uniform after clustering, which enables more robust computation of relevance required for the MRMR algorithm.

4.3 Robust computation of redundancy (RCR)

Class labels are not required for the computation of redundancy as shown in Eq.(4). Thus, large amounts of speech data without depression scores can be exploited for computing the redundancy. In the RCR approach, we propose using such unlabeled speech databases to compute redundancy for feature selection.

Figure 4 shows the distribution of correlations between features. Enriching the English database with unlabeled data had a significant effect on the distribution with a

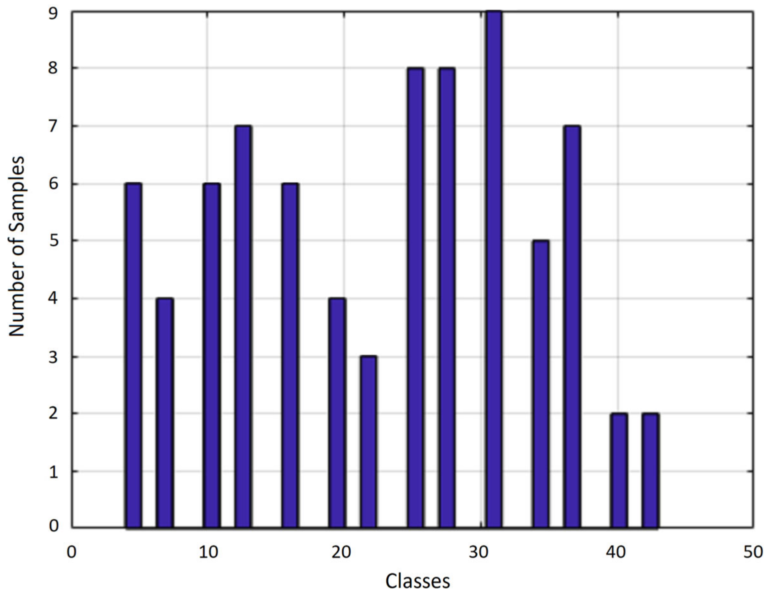

Fig. 3 Number of samples for each class after the clustering algorithm is applied. Turkish database is used

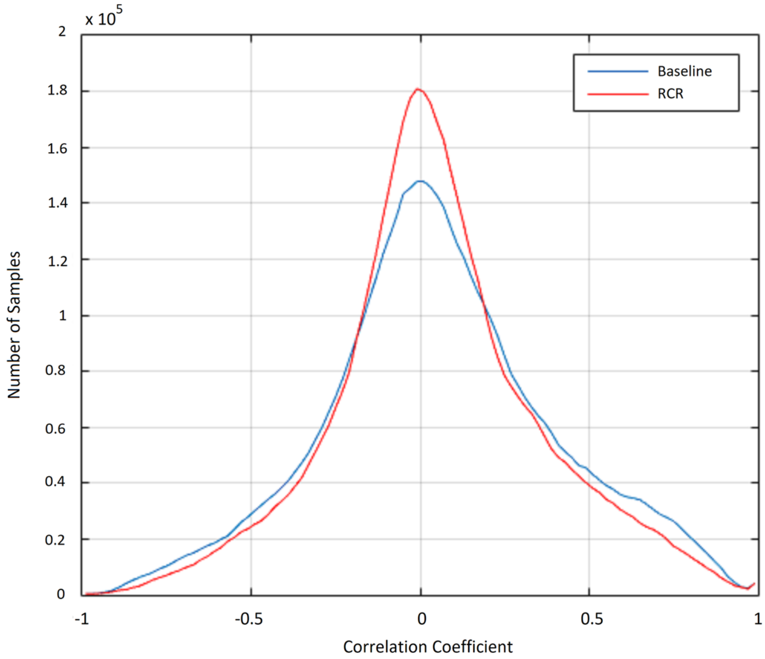

Fig. 4 Comparison of Pearson's correlation coefficient distributions for the English database with (RCR algorithm) and without (baseline) using the unlabelled data for computing correlations

sharper peak around zero and slightly suppressed tails. Thus, the distribution has lower variance after applying the RCR approach, which is expected to improve the feature selection performance.

5 Fusion with text-based features

5.1 Description of text-based features

Sentiments in questions and patient responses in the Turkish database were manually classified as positive, negative, and neutral. Examples of questions and answers with their sentiment tags are shown in Table 1. Feature vectors were generated from the sentiment tags where each dimension holds the frequency of question-answer sentiment pairs. Because there are three sentiments for questions and three sentiments for answers, a total of 9-dimensional sentiment feature vector was generated for each conversation.

Speech characteristics such as rate of speech and duration of responses can also be informative in depression studies. For example, given two positive responses from the subject, longer ones with elaboration are preferable to short ones. Similarly, short negative answers may indicate deeper depression than longer complaints. Thus, for each sentiment type, average rate of speech and average duration of responses were extracted using the timing information in the transcriptions. Because those two features were derived for each of the three sentiment types, 6-dimensional features were obtained for each conversation. Concatenating them with the 9 features described above, a total of 15 features were derived from the transcriptions. A summary of those features are shown in Table 2.

Table 1 Example of an interview in the Turkish database. Sentiment tags of both questions and answers are shown

	Phrase	Sentiment
Question:	Can you tell us a happy moment lately?	Positive
Answer:	I don't have one for a long time.	Negative
Question:	Can you tell us an unhappy moment lately?	Negative
Answer:	Everything goes well lately.	Positive
Question:	What is your favorite food?	Neutral
Answer:	I like stuffed peppers.	Neutral

Table 2 Descriptions of the text features derived from the transcripts of the conversations

Feature	Description
Average length of the utterances	Average length of subjects' negative, positive, and neutral answers separately. Three-dimensional feature.
Rate of speech	Rate of speech for negative, positive, and neutral answers separately. Three-dimensional feature.
Sentiment features	Sentiments of the question-answer pairs. All possible combinations sentiments are considered. Nine-dimensional feature.

5.2 Fusion of acoustic- and text-based features

The fusion algorithm is designed based on the observation that acoustics-only system sometimes makes large errors particularly when the subjects are very depressed or not depressed as shown in Fig. 5. Those large errors significantly impact the overall performance of the system and reduce its reliability.

In our approach, instead of using a typical score or feature fusion method, we propose a novel algorithm to adjust the acoustic-based scores using the text-based scores. In this approach, the data is first divided into two classes. Patients with BDI-II scores above 30 are tagged as class 1 and patients with scores below 18 are tagged as class 2.

If the acoustic-only system generates a depression level estimate that is above 30 or below 18 and if the text-only system also produces a score in the same range (agreement case), then the score from the acoustics-only system is used. If they are in disagreement, i.e., one of the systems produces an estimate that is in class 1 and the other produces an estimate that is in class 2, the final estimate is computed by fine-tuning the acoustics-only prediction by getting it closer to the opposite class.

In the case of disagreement, the following algorithm is used to adjust the estimate produced by the acoustics-based system. If the prediction of the acoustics-based system is p_{acou}, final prediction p_{final} is computed by the linear model:

$$p_{final} = \begin{cases} \alpha p_{acou} + \Gamma, & \text{if } p_{acou} < 18 \\ \alpha p_{acou} - \Gamma, & \text{if } p_{acou} > 30 \end{cases} \tag{12}$$

where α and Γ are constant parameters. Because the training data is limited, to avoid overfit, linear regression parameters were learned using a maximum a posteriori (MAP) approach where the prior distribution of α was modeled with a Gaussian distribution

$$p(\alpha) = \frac{1}{\sqrt{2\pi}} e^{-0.5(\alpha-1)^2} \tag{13}$$

where variance and mean were both set to 1. Mean is set to 1 so that p_{final} does not deviate significantly from p_{acou}. The prior distribution of Γ was also modelled with the Gaussian

$$p(\Gamma) = \frac{1}{\sqrt{2\pi}} e^{-0.5(\Gamma-\mu_g)^2} \tag{14}$$

where variance is set to 1 and mean is set to μ_g. Mean of the hyper-parameter Γ is learned from the data by setting α to 1 and learning the optimal Γ using leave-one-out.

6 Experimental setup

6.1 Databases

Three speech databases that are in Turkish, German, and English were used in this study. The databases are described below.

Turkish database: The Turkish database was collected at a hospital in Istanbul. It consists of 70 subjects. The mean age of the patients is 34. Fourteen of them are males and 56 of them are females. Beck scores of all subjects are available using the depression questionnaire the Beck Depression Inventory-II (BDI-II) [42]. Average BDI-II score of the patients is 23.45 with a standard deviation of 11.01.

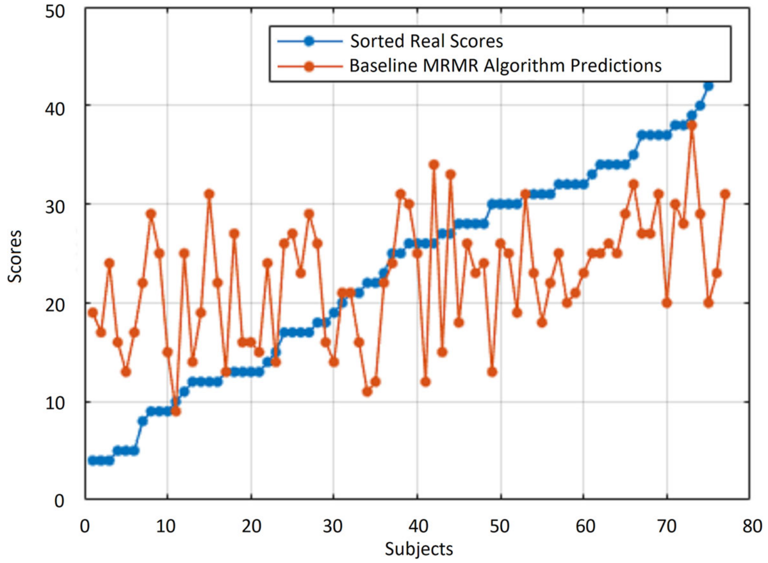

Fig. 5 Sorted real scores and the baseline MRMR predictions are compared using the Turkish and English databases for the regression task with the Turkish dataset

The Turkish database consists of interviews with the patients. Three types of questions were directed to the patients: neutral, positive, and negative questions. Each question type refers to the sentiment that we expect to invoke in the patient. Sentiments of the responses from the patients were manually tagged by three independent evaluators. Majority voting was used for the final sentiment label of each response. Examples of sentiment labels for the questions and answers are shown in Table 1.

Interviews consist of 16 questions. Mean length of the conversations is approximately 5 min. Total length of the recordings is 6 h. Recordings were done using a headphone microphone connected to the built-in sound card of a laptop with a sampling rate of 48 kHz.

German database: The German database, distributed as part of the AVEC 2014 challenge [43], consists of conversations with 84 patients. Some of the patients have multiple recordings with a period of 2 weeks. Even though Beck scores of the 100 recordings in the training and development data are available, the scores of the 50 recordings in the test data are not available. The mean age of German database subjects is 31.5. The duration of the recordings ranges from 6 s to 4 min. All recordings below 20 s were removed from the experiments, which left 98 recordings for processing.

English database: The English database is part of The Distress Analysis Interview Corpus (DAIC) [44]. It contains clinical interviews designed to help diagnose psychological distress conditions such as anxiety, depression, or post-traumatic stress disorder. The depression part of the corpus is the Wizard-of-Oz interviews that are conducted by a virtual interviewer. The depression scores of the patients were calculated by using the PHQ-8 depression inventory [45], which differs from the German and Turkish databases. The average depression severity of the training and development data is 6.67, and the standard deviation is 5.75. Total of 189 recordings from 189 patients is available.

6.2 Depression scores

We performed both regression and classification experiments in this study. For the classification task, the scores were split into two classes. For the BDI-II scores that were available in the Turkish and German databases, subjects that have scores below 18 were classified as non-depressed and other patients were classified as depressed. For the PHQ-8 scores available in the English database, subjects that have scores below 10 were classified as non-depressed and other patients were classified as depressed.

For regression, the ml-MRMR algorithm requires databases to use the same depression scale for computing within class statistics. However, in our experimental setup, the English database has PHQ-8 scores that range from 0 to 24 and the German and Turkish databases have Beck scores ranging from 0 to 63. Thus, a mapping function between those two scales was needed to carry out the multi-lingual regression experiments.

The BDI-II and the PHQ-8 are both widely used as self-rating scales to measure depression symptoms and severity of depression in psychiatric and normal populations [46]. Recall period for items for each scale is the last 2 weeks. There are 21 items in BDI-II and 8 items in PHQ-8. For PHQ-8, each item is scored on a four point scale (0–3) where 0 corresponds to not at all and 3 corresponds to nearly everyday. BDI-II items also have four point scales (0–3), but those do not measure the frequency of occurrence but rather general presence of a feeling/behavior.

The BDI-II was designed to correspond to Diagnostic and Statistical Manual of Mental Disorders, Fourth Edition (DSM-IV) criteria for diagnosing depressive disorders and includes items measuring cognitive, affective, somatic, and vegetative symptoms of depression. Similarly, PHQ-8 consist of the 8 criteria of DSM-IV and covers all of the DSM-IV criteria except self-harm.

Even though they have differences, the PHQ-8 and BDI-II scores are strongly correlated [47]. For PHQ-8, scores of 5, 10, 15, and 20 are cut-off points for mild, moderate, moderately severe, and severe depression respectively [45, 48]. For the BDI-II, the cut-off points for mild, moderate, and severe depression are 14, 20, and 29 respectively. Thus, the cut-off scores of the two measures have an approximately linear relationship.

Considering the strong correlation between the BDI-II and PHQ-8 scores, we mapped a given BDI-II score (s_b) to the corresponding PHQ-8 score, (s_p), by rounding $(24s_b)/63$ to nearest integer.

6.3 Acoustic features extraction

The open-source toolkit OpenSMILE [49] was used for acoustic feature extraction. The AVEC 2013 [43] and GeMAPS [15] feature extraction protocols were used. Feature vectors for AVEC 2013 include 32 energy- and spectral-related low-level descriptors (LLDs) and their functionals such as statistical functionals (maximum, mean, skewness, flatness, etc.), regression functionals (linear regression slope, quadratic regression coefficient a, etc.) and local minima/maxima-related functionals (mean and standard deviation of rising and falling slopes, etc.). 2268 dimensional features were extracted per speaker. Functionals were computed over 20 s time windows and averaged over the recording.

GeMAPS [15] has 18 low-level descriptors. Only the first 4 MFCC features are used in GeMAPS because those are more crucial for affect and paralinguistic voice analysis studies [15]. In addition, jitter, shimmer, loudness, and spectral slope were used. Similar to AVEC 2013, functionals of those low-level descriptors were also computed. The dimensionality of the final feature set is 62. Because GeMAPS is a hand-crafted feature set with reduced dimensionality, it is used for comparison with the proposed feature selection techniques here.

6.4 Baseline system

In the baseline system, MRMR feature selection method was first applied [50] to reduce the number of acoustic features. Support Vector Regression (SVR) was used for regression and SVMs were used for classification. Because the amount of training data is small, leave-one-out method was used for the Turkish and German experiments. For the English tasks, the training set has 107 subjects and the test set has 35 subjects. Because there is enough data both for training and test, leave-one-out method was not used for the English tasks.

The evaluation criteria for all regression experiments were root mean square error (RMSE), which is also used in the AVEC challenges [3, 43, 51, 52]. Statistical significance of the results were tested using the t test with $p < 0.05$.

The evaluation criteria for all classification experiments were F1-score, precision, and recall for both depressed and non-depressed subjects. For the classification tasks, statistical significance of results were measured with McNemar's test with $p < 0.05$.

7 Results and discussion

Two sets of experiments were conducted. In the first set, the proposed feature selection algorithms were tested and compared with the baseline MRMR algorithm for the German, Turkish, and English regression and classification tasks. The RCR algorithm proposed for redundancy computation in Section 4.3 was used only for the German and English tasks since unlabeled data is not available in the Turkish database. In the second test set, text-based features wee extracted and fused with the acoustic features for the Turkish database. The second set was performed only for the Turkish database because the transcriptions were not available for the German database; and, for the English database, the interviews were not in the question/answer format but rather a free-form talk between a human and computer.

7.1 Performance of the ml-MRMR feature selection and clustering algorithms

7.1.1 Turkish task

Regression: Table 3 shows the regression test results with the baseline MRMR and the ml-MRMR algorithms for the Turkish task. Lowest RMSE was 9.36 with the ml-MRMR ($N_{min} = 3$) algorithm using the Turkish-English data, and the improvement compared to the baseline was statistically significant. Similarly, ml-MRMR algorithm using the Turkish-German data outperformed the baseline system, and the difference was statistically significant. Moreover, the ml-MRMR algorithms performed better than the Gemaps feature set. Performance was better when N_{min} was set to 3 compared to setting it to 5.

Table 3 Regression performance of the baseline MRMR and ml-MRMR methods for the Turkish task when the minimum occurrence threshold (N_{min}) is sets 3 and 5. In the underlined bold case, improvement is significant compared to the baseline system. The result with Gemaps feature set was 11.48

Dim	Baseline	$N_{min}(3)$	$N_{min}(5)$	$N_{min}(3)$	$N_{min}(5)$
		(Tr+Ger)	(Tr+Ger)	(Tr+Eng)	(Tr+Eng)
2	13.30	10.84	11.98	10.79	10.88
3	12.30	**10.51**	11.26	10.40	11.94
4	12.45	10.85	**10.74**	9.85	12.68
5	12.56	10.58	11.23	**9.36**	13.65
10	12.45	10.82	12.13	11.87	13.93
15	12.08	11.12	12.00	11.99	13.30
20	12.87	11.91	11.46	10.92	12.00
40	13.28	12.67	11.98	10.93	10.31
80	11.58	12.28	13.06	10.80	10.50
100	11.75	11.95	13.08	10.88	**10.23**
200	11.32	11.55	12.14	11.05	11.06
400	11.42	11.72	12.00	10.99	11.23
800	11.31	11.39	11.35	11.10	11.08

For regression, the clustering algorithm described in Section 4.2 was used with 2, 9, and 15 clusters instead of the 45 distinct classes available in the Beck scores. Results are shown in Table 4. Even though the system with 15 clusters significantly outperformed the baseline system, the improvement was still below what was obtained with the multi-lingual MRMR approach.

Table 4 Regression results with feature selection using the clustering approach with 2, 9, and 15 clusters. Turkish database is used. Statistically significant ($p < 0.05$) improvement is shown in underlined bold

Dim	Baseline	2-Cluster	9-Cluster	15-Cluster
5	12.56	11.35	13.14	11.99
10	12.45	10.95	13.42	12.25
15	12.08	11.13	13.07	11.75
20	12.87	11.74	13.23	12.95
40	13.28	12.33	13.73	12.06
80	11.58	12.72	13.33	**10.83**
100	11.75	13.22	13.09	10.97
200	11.32	11.72	12.66	11.50
400	11.42	11.83	12.00	11.40
800	11.31	11.62	11.70	11.64

Classification: The ml-MRMR algorithm was not applied directly in the case of classification because there are only two classes and each class has enough number of samples (27 non-depressed and 50 depressed subjects). However, it is still possible to use the ml-MRMR algorithm in the binary classification case by populating each class from a cross-lingual dataset before dividing the data into two classes. After each regression class (1 to 45) is populated with the cross-lingual samples, training data is split into two classes for the classification task.

Classification results are shown in Table 5. Even though ml-MRMR algorithm improves the performance, the improvement was not found to be statistically significant. Thus, in the classification case, ml-MRMR algorithm was not as effective because enough Turkish data was available in each class. The system trained with only the text-based features significantly outperformed the other systems.

7.1.2 English task

Regression: Table 6 shows the regression results for the English task. Best result was obtained by using the ml-MRMR algorithm with Turkish ($N_{min} = 5$). Even though ml-MRMR using the German database performed better than the baseline, the improvement was not significant. Note that the English database uses the PHQ-8 scores that are coarser than the Beck scores used in the German and Turkish databases. Thus, there are more samples for each class and using ml-MRMR algorithm with $N_{min} = 3$ was not possible for the English case.

Classification: Table 7 shows the classification results for the English task. Similar to the regression task, the ml-MRMR algorithm with Turkish using $N_{min} = 5$ outperformed the baseline system, and, when German data was used, performance did not significantly change. The ml-MRMR algorithm using the English-Turkish datasets improved the F1 scores of both depressed and non-depressed subjects. Improvement for the depressed subjects were higher compared to the non-depressed subjects.

7.1.3 German task

Since the German dataset contains unlabeled data, RCR algorithm was used to compute feature correlations in addition to the ml-MRMR algorithms. Results using those two algorithms for the regression and classification tasks are discussed below.

Table 5 Best classification results for Turkish Task. Avec 2013 feature set used for all results except Gemaps row. There are 50 depressed and 27 non-depressed subjects in the database. Number of selected features are shown in parenthesis for each case. Results are statistically insignificant except the classification using text-only features

Method	Classes	Precision	Recall	F1-score
Baseline MRMR(3)	**Non-depressed**	0.61	0.40	0.48
	Depressed	0.72	0.86	0.78
	Average	0.67	0.63	0.63
Tr+Eng $N_{min} = 3$ ***ml-MRMR(40)***	**Non-depressed**	0.58	0.52	0.55
	Depressed	0.75	0.80	0.78
	Average	0.67	0.66	0.66
Tr+Eng $N_{min} = 5$ ***ml-MRMR(400)***	**Non-depressed**	0.59	0.48	0.53
	Depressed	0.74	0.82	0.78
	Average	0.66	0.65	0.66
Tr+Ger $N_{min} = 3$ ***ml-MRMR(100)***	**Non-depressed**	0.42	0.29	0.34
	Depressed	0.67	0.78	0.72
	Average	0.54	0.54	0.53
Tr+Ger $N_{min} = 5$ ***ml-MRMR(15)***	**Non-depressed**	0.47	0.52	0.49
	Depressed	0.72	0.68	0.70
	Average	0.60	0.60	0.60
Only text features MRMR(7)	**Non-depressed**	0.78	0.40	0.53
	Depressed	0.74	0.94	**0.83**
	Average	0.76	0.67	**0.68**
GEMAPS	**Non-depressed**	0.38	0.37	0.37
	Depressed	0.66	0.68	0.67
	Average	0.52	0.53	0.52

Table 6 Regression performance of the ml-MRMR method for the English task when the minimum occurrence threshold (N_{min}) is set 5. In the underlined bold case, improvement is significant compared to the baseline system. Best result with the Tr+Eng also significantly outperformed the Gemaps feature set that had an RMSE of 6.72

Dim	Baseline	$N_{min} = 5$ (Ger+Eng)	$N_{min} = 5$ (Tr+Eng)
3	6.85	6.66	6.66
4	6.87	7.21	7.20
5	7.31	7.78	7.78
10	7.85	7.05	7.05
15	8.08	7.94	7.70
20	7.89	7.63	7.46
40	7.25	7.12	6.63
80	7.67	**6.26**	6.37
100	7.57	6.38	**6.15**
200	7.13	6.42	6.70
400	6.95	6.72	6.73
800	6.92	6.74	6.66

Regression: Regression performance of the baseline and the proposed ml-MRMR (with Turkish) and RCR feature selection algorithms for the German task are shown in Table 8. ml-MRMR algorithm was not effective when German was used with English. When German was used with Turkish, performance improved for $N_{min} = 3$; however, the improvement was not significant. Improvement with $N_{min} = 5$ was found to be significant only when the RCR algorithm was also used. RCR algorithm was not effective when it was used without ml-MRMR.

Table 7 Best classification results for the development set of the English task. Multi-lingual methods annotated with ml. There are 23 non-depressed and 12 depressed subjects in the database. Avec 2013 feature set used for all results except Gemaps tab. Number of selected features are shown in parenthesis for each case. Improvement with ml-MRMR using the Turkish database is statistically significant compared to the baseline MRMR algorithm

Method	Classes	Precision	Recall	F1-score
Baseline ***MRMR(20)***	**Non-depressed**	0.76	0.96	0.85
	Depressed	0.83	0.42	0.56
	Average	0.79	0.69	0.71
Eng+Tr $N_{min} = 5$ ***ml-MRMR(100)***	**Non-depressed**	0.79	1	**0.88**
	Depressed	1	0.50	**0.66**
	Average	0.89	0.75	**0.77**
Eng+Ger $N_{min} = 5$ ***ml-MRMR(20)***	**Non-depressed**	0.76	0.96	0.85
	Depressed	0.83	0.42	0.56
	Average	0.79	0.69	0.71
GEMAPS	**Non-depressed**	0.70	0.82	0.76
	Depressed	0.50	0.33	0.40
	Average	0.60	0.58	0.58

Classification: Classification performance of the baseline and the proposed ml-MRMR and RCR algorithms for the German task are shown in Table 9. The ml-MRMR algorithm with Turkish using $N_{min} = 3$ outperformed the rest of the systems when RCR was used.

7.2 Performance of score fusion

Text-based features were only available for the Turkish dataset. Therefore, results for the score fusion algorithm are reported only for the Turkish dataset.

7.2.1 Regression task

Table 10 shows results when speech-based features were fused with text-based features using the proposed approach described in Section 5.2. Fusion algorithm significantly improved the performance (p value=0.00006) compared to the baseline case by reducing the error by more than 25% using the ml-MRMR algorithm with English and Turkish ($N_{min} = 3$) datasets. Spread of the prediction errors is substantially reduced after fusion as shown in Fig. 6.

Comparison of the real and predicted scores are shown in Figs. 5 and 7 for the baseline and the best ml-MRMR algorithm with fusion. Predictions get closer to the true scores and errors significantly decrease with the proposed fusion method, which can be seen when Figs. 4 and 7 are compared. The best RMSE is 8.30, which is interestingly obtained with only 5 features.

Three of the 5 selected acoustic features are MFCC related: peak standard deviation of MFFC-5, amplitude mean of maxima for MFCC-5 and mean segment length of MFCC-14. The other two is mean of the rising slope for spectral harmonicity and up-level time (25) of spectral flatness.

The ml-MRMR algorithm with German and Turkish datasets ($N_{min} = 3$) worked well compared to the baseline as shown in the fourth column in Table 10. Still, it did not perform as well as the Turkish and English case. Moreover, its performance was not significantly different from the base-fusion. These results are in agreement with the results reported in Table 3 where performance with Turkish and English datasets was better compared to the Turkish and German datasets.

Sixth column in Table 10 shows the results obtained with the clustering approach together with the fusion method. That algorithm not only outperformed the baseline but also significantly outperformed the base-fusion algorithm. However, it performed worse than the best performing ml-MRMR system.

Table 8 Regression performance of the ml-MRMR Methods for the German database when the minimum occurrence threshold N_{min} is sets 3 and 5. Results are shown both when the RCR algorithm is used and not used. In the underlined bold case, improvement is significant compared to the baseline system. Result with Gemaps feature set was: 10.14

Dim	Baseline	$N_{min} = 5$	$N_{min} = 5$ and RCR	$N_{min} = 3$	$N_{min} = 3$ and RCR	RCR
		(Ger+Tr)	(Ger+Tr)	(Ger+Tr)	(Ger+Tr)	
10	9.90	9.97	9.99	10.37	12.39	10.02
15	9.81	10.21	**9.43**	10.13	12.12	10.08
20	9.86	10.32	9.52	9.84	10.68	9.74
40	10.25	10.35	10.45	9.73	11.36	10.22
80	10.69	9.93	9.88	**9.42**	10.93	10.06
100	10.48	9.93	9.74	9.50	10.54	10.17
200	10.12	10.00	10.38	9.69	10.28	10.44
400	10.14	9.79	10.21	9.58	10.29	10.13
800	10.08	9.86	10.11	9.91	10.11	9.89
1000	10.02	9.85	10.14	9.79	10.16	9.98

7.2.2 Classification task

For the classification task, text-only model predictions outperformed the acoustic system predictions as shown in Table 5. When the text-only model predictions were fused with the acoustic predictions, the F1-scores outperformed both modalities when English ($N_{min} = 3$) was used to supplement the Turkish samples as shown in Table11. The result is statistically significant with p value of $p = 0.02$.

When German ($N_{min} = 3$) was used with Turkish, the F1-score for the depressed case slightly improved. However, the F1-score for the non-depressed case did not improve. Performance of the Turkish-German and Turkish-English systems were not significantly different for the classification task.

Table 9 Classification results for German Task. ml-MRMR was used with RCR. There are 56 non-depressed and 44 depressed subjects in the database. Best results are shown in bold. Number of selected features are shown in parenthesis for each case. The improvements are not statistically significant with respect to baseline system

Method	Classes	Precision	Recall	F1-score
Baseline MRMR(10)	**Non-depressed**	0.76	0.87	0.81
	Depressed	0.80	0.65	0.72
	Average	0.78	0.76	0.77
Ger+Tr $N_{min} = 3$ ***ml-MRMR(40)***	**Non-depressed**	0.83	0.79	**0.81**
	Depressed	0.74	0.80	**0.77**
	Average	0.79	0.80	**0.79**
Ger+Eng $N_{min} = 5$ ***ml-MRMR(800)***	**Non-depressed**	0.73	0.78	0.75
	Depressed	0.70	0.63	0.66
	Average	0.72	0.71	0.71
GEMAPS	**Non-depressed**	0.74	0.70	0.72
	Depressed	0.64	0.68	0.66
	Average	0.69	0.69	0.69

8 Discussion

The ml-MRMR algorithm was the best performing feature selection algorithm in regression tasks. Populating the Turkish dataset with English, English dataset with Turkish, and German dataset with Turkish generated the best results. The ml-MRMR algorithm was effective for German only when it was used together with the RCR algorithm. Thus, the cross-lingual population of depression classes was not as effective with German as the other two languages. The acoustic characteristics of Turkish and English seems to be closer to each other and those two languages complement each other better than they do with German.

Note that the English and German datasets have significantly more training data compared to the Turkish dataset. Thus, because the Turkish dataset is more sparse across depression classes, $N_{min} = 3$ performed better than $N_{min} = 5$. Using $N_{min} > 5$ caused overly aggressive population of the Turkish dataset with cross-lingual data, which degraded the performance. Similarly, the English and German datasets performed better when $N_{min} = 5$ since using a lower N_{min} caused insufficient population of their depression classes with cross-lingual data.

The proposed feature selection algorithms were designed for cases when the number of classes are large. Thus, they were not as effective for binary classification tasks as they were in regression tasks. Still, some improvement was observed for the English classification task when ml-MRMR was used with Turkish.

The text features were tested for the Turkish dataset and found to outperform all acoustic feature sets for

Table 10 Regression results after fusing with text classification. Turkish database was used since transcriptions are not available in the required for the other databases. Baseline acoustic system predictions are used in base-fusion. Bold results show cases where the improvement is significant compared to the baseline case but not to the base-fusion case. In the underlined bold case, improvement is significant both compared to the baseline system and the base-fusion system

Dim	Baseline	Base-fusion	Fusion Tr+Ger	Fusion Tr+Eng	Fusion
			$N_{min} = 3$	$N_{min} = 3$	15 Clus.
3	12.30	10.71	9.76	9.43	9.75
4	12.45	10.68	**9.54**	8.88	9.74
5	12.56	10.69	9.61	**8.30**	9.78
10	12.45	10.91	9.83	9.74	10.05
15	12.08	10.44	9.63	10.12	9.95
20	12.87	10.91	9.65	9.62	**9.66**
40	13.28	11.33	10.45	9.55	10.88
80	11.58	10.12	10.67	9.37	9.99
100	11.75	10.25	10.77	9.45	10.24
200	11.32	**10.03**	9.98	9.76	10.40
400	11.42	10.29	10.00	9.60	9.78
800	11.31	10.31	10.19	9.79	9.88

the classification task. Thus, the sentiment-based text features were found to be effective for binary classification of depression. Similarly, when the text features were fused with the acoustic features using the proposed fusion algorithm, performance with the Turkish dataset significantly improved both for regression and classification tasks. Fusion of text and acoustic features outperformed both of the feature sets.

The clustering algorithm helped improve the performance of the Turkish dataset with and without fusion of text features. However, it was not as effective as the multi-lingual approach. Thus, cross-lingual population of

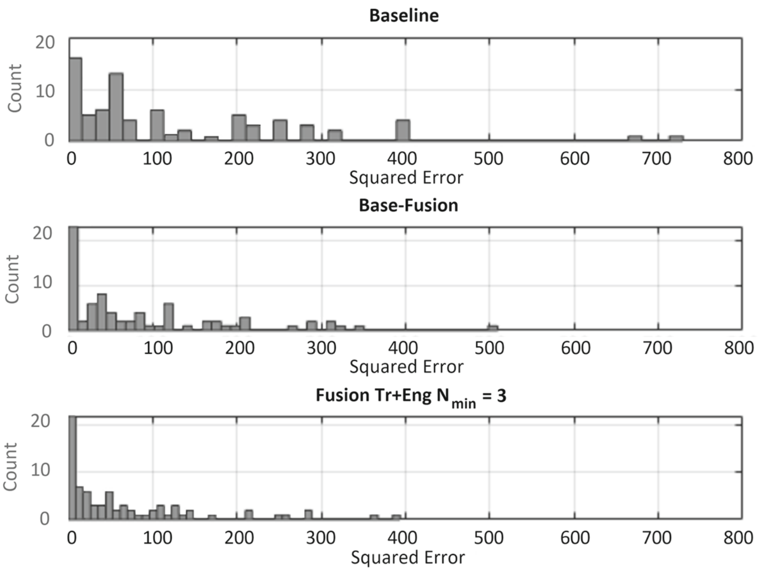

Fig. 6 Distribution of squared errors for the baseline MRMR case is shown in the top figure for the Turkish task. The middle figure shows the squared error distribution for the baseline MRMR case after fusion. Bottom figure shows the squared error distribution of the ml-MRMR system with $N_{min} = 3$ with English and after fusion

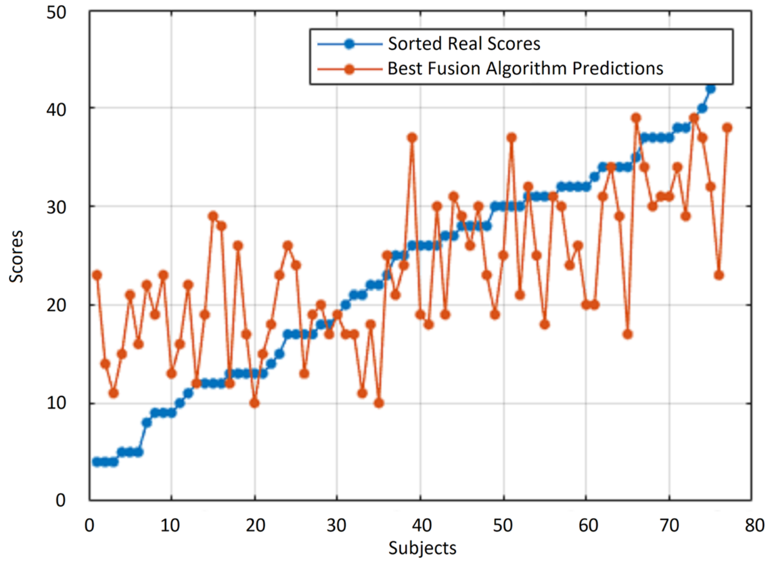

Fig. 7 Sorted real scores and predictions with the ml-MRMR ($N_{min} = 3$) are compared using the Turkish and English databases for the regression task with the Turkish dataset

depression classes was found to be more effective than simply reducing the number of classes through clustering. The multi-lingual approach allows computation of relevance with more data per class without reducing the resolution of the depression scale. If the languages have similar acoustic representations of depression, such as Turkish and English as found in our experiments, then the multi-lingual approach outperforms the within-language clustering algorithm.

Table 11 Classification results after fusing acoustic- and text-based classifier outputs. Turkish database was used. The results are statistically insignificant except $N_{min} = 3$ with German

Method	Classes	Precision	Recall	F1-score
	Non-depressed	0.61	0.40	0.48
Baseline	**Depressed**	0.72	0.86	0.78
MRMR(3)	**Average**	0.67	0.63	0.63
	Non-depressed	0.71	0.63	**0.67**
Fusion Tr+Eng	**Depressed**	0.81	0.86	0.83
$N_{min} = 3$	**Average**	0.76	0.75	**0.75**
ml-MRMR(100)	**Non-depressed**	0.75	0.55	0.63
Fusion Tr+Ger	**Depressed**	0.78	0.90	**0.84**
$N_{min} = 3$	**Average**	0.76	0.73	0.73
ml-MRMR(4)	**Non-depressed**	0.78	0.40	0.53
Only Text	**Depressed**	0.74	0.94	0.83
Features MRMR(7)	**Average**	0.76	0.67	0.68

8.1 Common selected features among languages

In this study, we explored the features that are most effective at predicting depression for three different languages. In addition, we did further analysis of our results to find features that are common across those three languages. Table 12 shows overlapping features between Turkish-English, Turkish-German, and English-German pairs within the top 150 MRMR-selected features. Overlapping features and their functionals are described in Tables 13 and 14, respectively.

Features that are based on spectral harmonicity and energy in 1000–4000 Hz are dominant in the Turkish-English comparison. 1000–4000 Hz typically contain the second and the third formants. Thus, the rate of change of those formants, measured with up-level time, and distance between them appear to be strong indicators of depression for Turkish and English. Similarly, change in spectral harmonicity is also a strong indicator for both languages.

Interestingly, MFCC features are dominant in the Turkish-German comparison. MFCC features are related to the envelope of the spectrum. Thus, changes in the locations of all formants and their bandwidths during speech are important indicators of depression detection both for Turkish and German.

Overlapping features between English and German have a mix of energy, spectral harmonicity, jitter, and MFCC features. As opposed to Turkish-English comparison where the second and third formants are important, energy in the 250–650 Hz that typically contains the first

Table 12 Overlapping features in language pairs that are in the top 150 MRMR-selected features

Language combination	LLD & functionals
Turkish-English	MFCC-13 - Relative mean of peaks
	Energy in Band 1000–4000 Hz - Kurtosis
	Spectral harmonicity - Up-level Time:90
	Spectral harmonicity - Up-level Time:50
Turkish-German	Energy in band 1000–4000 Hz - Up-level Time:25
	Energy in band 1000–4000 Hz - Relative Mean of Peaks
	Energy in band 1000–4000 Hz - Minimum
	Segment length
	Spectral skewness - IQR 2-3
	Spectral skewness - IQR 1-3
	MFCC-4 - Mean distance between peaks
	MFCC-9 - Rise-time
	MFCC-11 - Rise-time
	MFCC-13 - Skewness
	MFCC-14 - Flatness
	MFCC-14 - Kurtosis
English-German	Energy in band 250–650 Hz - Quartile 1
	Energy in band 250–650 Hz - Percentile 1.0
	Spectral harmonicity - Up-level time:25
	JitterLocal - Mean
	JitterLocal - Standard deviation
	MFCC-10 - Mean segment length
	MFCC-14 - Up-level time:90
	MFCC-16 - Mean distance between peaks
	JitterDDP - Up-level time:50
	JitterDDP - Up-level time:90

Table 13 Overlapping low-level descriptors in language pairs that are in the top 150 MRMR-selected features

Low-level descriptor	Description
MFCC 1-16	Mel Frequency Cepstral Coefficient is a common used automatic speech recognition (ASR) feature, in the Avec 2013 feature set 16 dimension were used.
Energy	Sum squares of amplitudes of a signal.
Spectral harmonicity	Number of the harmonics in a signal.
Spectral skewness	The third order moment of the power spectrum.
Jitter (local)	Variation of the fundamental period from one single period towards the next.
Jitter (DDP)	Delta period-to-period jitter can be defined as “Jitter of the Jitter”. It is explained as the change between two successive period-to-period jitters.

Table 14 Overlapping functionals in language pairs that are in the top 150 MRMR-selected features

Statistical functionals	Description
Relative mean of peaks	Proportion of the mean of the peak amplitudes to the mean of windowed feature.
Kurtosis	Fourth order moment.
Skewness	Third order moment.
Up-level time	Number of frames that the feature is above a threshold. The threshold percentiles are set to 25, 50, 75, and 90.
Minimum segment length	Minimum length of a particular segment.
Mean segment length	Arithmetic mean of a particular segment.
Inter quartile range 1-2-3 (IQR)	The range between two percentiles. The possible combination of quartiles are 1–3, 1–2, and 2–3.
Mean distance between peaks	The mean of distances between the peaks.
Rise-time	The time where the feature contour is rising.
Percentile 1.0	The minimum value of a feature.

formant appears to be important and overlapping for English and German. Jitter, which quantifies pitch variations is also important both for English and German but not for Turkish.

9 Conclusion and future work

We investigated exploiting multi-lingual databases for feature selection in the context of depression assessment. Proposed algorithms were effective especially for the regression tasks where there is limited amounts of data for each class. As a second contribution, we proposed novel features derived from transcriptions and fused them with the acoustic features, which significantly improved the performance.

The results are significant because they indicate that there are similarities between entirely different languages in the way that they manifest depression. Thus, our findings is a step towards using larger multi-lingual databases for depression detection.

The focus of this work was multi-lingual feature selection algorithms and not the classification algorithms. Thus, even though the SVM and SVR algorithms are solid baselines when the amount of training data is limited, in future work, we will experiment with other types of classification/regression algorithms such as gradient boosting and random forests.

Even though the Turkish database used here is unique to this work, the English and German databases are publicly available and have been used together in the literature as discussed in Section 2.3. A comparison of our proposed techniques with the previously proposed multilingual techniques will be done in future work. Moreover, because our method is focused on feature selection, it can also be used together with the previously proposed methods, which will be investigated in future work.

Another natural extension of the proposed algorithms is to add more languages to our database and continue to improve and analyze the feature selection process, which will be done in future work. In that context, we believe that our text features are also language-independent and we will investigate fusion algorithms in a multi-lingual setting with more data, such as Arabic, collected in the format required by our text-based features.

Availability of the data and materials

The Turkish database is not available for public use.

Acknowledgements

We would like to thank to the Sisli Etfal Hospital for the help they provided for collecting the data. We would like to thank to the Ozyegin University for the provided means.

Authors' contributions

Dr. Selime Celik and Dr. Asli Besirli were responsible for collecting the data of speech and psychological information. Cenk Demiroglu was the advisor and the implementation of the base version of MRMR algorithm. The rest has been done by Yasin Ozkanca. The authors read and approved the final manuscript.

Funding

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Department of Electrical and Electronics Engineering, Ozyegin University, Istanbul, Turkey. [2]University of Health Sciences Turkey, Şişli Hamidiye Etfal Training and Research Hospital, Department of Psychiatry, Istanbul, Turkey.

Received: 4 February 2020 Accepted: 11 October 2020
Published online: 17 November 2020

References

1. A. Halfin, Depression: the benefits of early and appropriate treatment. Am. J. Manage Care. **13**, 92–7 (2007)
2. Depression and other common mental disorders: global health estimates. Geneva World Health Organ., 13 (2017)
3. M. Valstar, J. Gratch, B. Schuller, F. Ringeval, D. Lalanne, M. Torres Torres, S. Scherer, G. Stratou, R. Cowie, M. Pantic, in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. Avec 2016: Depression, mood, and emotion recognition workshop and challenge (Association for Computing Machinery, New York, 2016), pp. 3–10
4. J. C. Mundt, A. P. Vogel, D. E. Feltner, W. R. Lenderking, Vocal acoustic biomarkers of depression severity and treatment response. Biol. Psychiatry. **72**(7), 580–587 (2012)
5. D. J. France, R. G. Shiavi, S. Silverman, M. Silverman, M. Wilkes, Acoustical properties of speech as indicators of depression and suicidal risk. IEEE Trans. Biomed. Eng. **47**(7), 829–837 (2000)
6. B. Stasak, J. Epps, R. Goecke, in *Proc. Interspeech 2017*. Elicitation design for acoustic depression classification: an investigation of articulation effort, linguistic complexity, and word affect (International Speech Communication Association, France, 2017), pp. 834–838. https://doi.org/10.21437/Interspeech.2017-1223
7. N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, T. F. Quatieri, A review of depression and suicide risk assessment using speech analysis. Speech Comm. **71**, 10–49 (2015)
8. F. Or, J. Torous, J.-P. Onnela, High potential but limited evidence: using voice data from smartphones to monitor and diagnose mood disorders. Psychiatr. Rehabil. J. **40**(3), 320 (2017)
9. O. Simantiraki, P. Charonyktakis, A. Pampouchidou, M. Tsiknakis, M. Cooke, in *Proc. Interspeech 2017*. Glottal source features for automatic speech-based depression assessment (International Speech Communication Association, France, 2017), pp. 2700–2704. https://doi.org/10.21437/Interspeech.2017-1251
10. B. S. Helfer, T. F. Quatieri, J. R. Williamson, D. D. Mehta, R. Horwitz, B. Yu, in *Interspeech*. Classification of depression state based on articulatory precision (International Speech Communication Association, France, 2013), pp. 2172–2176
11. N. Cummins, V. Sethu, J. Epps, J. Krajewski, in *Interspeech*. Probabilistic acoustic volume analysis for speech affected by depression (International Speech Communication Association, France, 2014), pp. 1238–1242
12. B. Vlasenko, H. Sagha, N. Cummins, B. Schuller, in *Proc. Interspeech 2017*. Implementing gender-dependent vowel-level analysis for boosting speech-based depression recognition (International Speech Communication Association, France, 2017), pp. 3266–3270. https://doi.org/10.21437/Interspeech.2017-887
13. A. Afshan, J. Guo, S. J. Park, V. Ravi, J. Flint, A. Alwan. Effectiveness of voice quality features in detecting depression (International Speech Communication Association, France, 2018), pp. 1676–1680
14. F. Ringeval, B. Schuller, M. Valstar, N. Cummins, R. Cowie, L. Tavabi, M. Schmitt, S. Alisamir, S. Amiriparian, E.-M. Messner, *et al*, in *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*. Avec 2019 workshop and challenge: state-of-mind, detecting depression with ai, and cross-cultural affect recognition (Association for Computing Machinery, New York, 2019), pp. 3–12
15. F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, *et al*, The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. IEEE Trans. Affect. Comput. **7**(2), 190–202 (2016)
16. M. Schmitt, F. Ringeval, B. W. Schuller, in *Interspeech*. At the border of acoustics and linguistics: Bag-of-audio-words for the recognition of emotions in speech (International Speech Communication Association, France, 2016), pp. 495–499
17. F. Ringeval, B. Schuller, M. Valstar, R. Cowie, H. Kaya, M. Schmitt, S. Amiriparian, N. Cummins, D. Lalanne, A. Michaud, *et al*, in *Proceedings of the 2018 on Audio/visual Emotion Challenge and Workshop*. Avec 2018 workshop and challenge: bipolar disorder and cross-cultural affect recognition (Association for Computing Machinery, New York, 2018), pp. 3–13
18. S. Amiriparian, M. Gerczuk, S. Ottl, N. Cummins, M. Freitag, S. Pugachevskiy, A. Baird, B. W. Schuller, in *INTERSPEECH*. Snore sound classification using image-based deep spectrum features. vol. 434 (International Speech Communication Association, France, 2017), pp. 3512–3516
19. J. F. Cohn, T. S. Kruez, I. Matthews, Y. Yang, M. H. Nguyen, M. T. Padilla, F. Zhou, F. De la Torre, in *Affective Computing and Intelligent Interaction and Workshops, 2009. ACII 2009. 3rd International Conference On*. Detecting depression from facial actions and vocal prosody (IEEE Computer Society, Los Alamitos, 2009), pp. 1–7
20. M. Kächele, M. Glodek, D. Zharkov, S. Meudt, F. Schwenker, Fusion of audio-visual features using hierarchical classifier systems for the recognition of affective states and the state of depression. Depression. **1**(1), 671–678 (2014)
21. V. Jain, J. L. Crowley, A. K. Dey, A. Lux, in *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*. Depression estimation using audiovisual features and fisher vector encoding (Association for Computing Machinery, New York, 2014), pp. 87–91
22. R. Gupta, S. S. Narayanan, in *INTERSPEECH*. Predicting affective dimensions based on self assessed depression severity (International Speech Communication Association, France, 2016), pp. 1427–1431
23. R. Gupta, S. Sahu, C. Espy-Wilson, S. S. Narayanan, in *Proc. Interspeech 2017*. An affect prediction approach through depression severity parameter incorporation in neural networks (International Speech Communication Association, France, 2017), pp. 3122–3126. https://doi.org/10.21437/Interspeech.2017-120
24. J. R. Williamson, T. F. Quatieri, B. S. Helfer, G. Ciccarelli, D. D. Mehta, in *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*. Vocal and facial biomarkers of depression based on motor incoordination and timing (Association for Computing Machinery, New York, 2014), pp. 65–72

25. B.-Q. Li, L.-L. Hu, L. Chen, K.-Y. Feng, Y.-D. Cai, K.-C. Chou, Prediction of protein domain with MRMR feature selection and analysis. PLoS ONE. **7**(6), 39308 (2012)
26. Y. Cai, T. Huang, L. Hu, X. Shi, L. Xie, Y. Li, Prediction of lysine ubiquitination with MRMR feature selection and analysis. Amino Acids. **42**(4), 1387–1395 (2012)
27. M. Pal, G. M. Foody, Feature selection for classification of hyperspectral data by SVM. IEEE Trans. Geosci. Remote Sens. **48**(5), 2297–2307 (2010)
28. Y. Gong, C. Poellabauer, in *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*. Topic modeling based multi-modal depression detection (Association for Computing Machinery, New York, 2017), pp. 69–76
29. M. A. Hall, Correlation-based feature subset selection for machine learning. Thesis submitted in partial fulfillment of the requirements of the degree of Doctor of Philosophy at the University of Waikato (1998)
30. S. Alghowinem, R. Goecke, J. Epps, M. Wagner, J. Cohn, in *Interspeech 2016*. Cross-cultural depression recognition from vocal biomarkers (International Speech Communication Association, France, 2016), pp. 1943–1947
31. F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, *et al*, The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. IEEE Trans. Affect. Comput. **7**(2), 190–202 (2015)
32. R. Gupta, N. Malandrakis, B. Xiao, T. Guha, M. Van Segbroeck, M. Black, A. Potamianos, S. Narayanan, in *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*. Multimodal prediction of affective dimensions and depression in human-computer interactions (Association for Computing Machinery, New York, 2014), pp. 33–40
33. M. R. Morales, R. Levitan, in *Spoken Language Technology Workshop (SLT), 2016 IEEE*. Speech vs. text: a comparative analysis of features for depression detection systems (IEEE, 2016), pp. 136–143
34. S. Scherer, G. Stratou, J. Gratch, L.-P. Morency, in *Interspeech*. Investigating voice quality as a speaker-independent indicator of depression and PTSD (International Speech Communication Association, France, 2013), pp. 847–851
35. J. R. Williamson, E. Godoy, M. Cha, A. Schwarzentruber, P. Khorrami, Y. Gwon, H.-T. Kung, C. Dagli, T. F. Quatieri, in *Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge*. Detecting depression using vocal, facial and semantic communication cues (Association for Computing Machinery, New York, 2016), pp. 11–18
36. E.-M. Rathner, J. Djamali, Y. Terhorst, B. Schuller, N. Cummins, G. Salamon, C. Hunger-Schoppe, H. Baumeister, How did you like 2017? detection of language markers of depression and narcissism in personal narratives. Future. **1**(2.58), 0 (2018)
37. T. Al Hanai, M. M. Ghassemi, J. R. Glass, in *Interspeech*. Detecting Depression with Audio/Text Sequence Modeling of Interviews (International Speech Communication Association, France, 2018), pp. 1716–1720
38. L. Yang, H. Sahli, X. Xia, E. Pei, M. C. Oveneke, D. Jiang, in *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*. Hybrid depression classification and estimation from audio video and text information (Association for Computing Machinery, New York, 2017), pp. 45–51
39. M. Rodrigues Makiuchi, T. Warnita, K. Uto, K. Shinoda, in *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*. Multimodal fusion of bert-CNN and gated CNN representations for depression detection (Association for Computing Machinery, New York, 2019), pp. 55–63
40. V. Mitra, E. Shriberg, D. Vergyri, B. Knoth, R. M. Salomon, in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference On*. Cross-corpus depression prediction from speech (IEEE, 2015), pp. 4769–4773
41. J. Novikova, A. Balagopalan, in *QinAI Workshop at NeurIPS*. On Speech Datasets in Machine Learning for Healthcare, (Vancouver, 2019)
42. A. T. Beck, R. A. Steer, G. K. Brown, Beck depression inventory-ii. San Antonio. **78**(2), 490–8 (1996)
43. M. Valstar, B. Schuller, K. Smith, T. Almaev, F. Eyben, J. Krajewski, R. Cowie, M. Pantic, in *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*. Avec 2014: 3D dimensional affect and depression recognition challenge (Association for Computing Machinery, New York, 2014), pp. 3–10
44. J. Gratch, R. Artstein, G. M. Lucas, G. Stratou, S. Scherer, A. Nazarian, R. Wood, J. Boberg, D. DeVault, S. Marsella, *et al*, in *LREC*. The distress analysis interview corpus of human and computer interviews (European Language Resources Association (ELRA), 2014), pp. 3123–3128
45. K. Kroenke, T. W. Strine, R. L. Spitzer, J. B. Williams, J. T. Berry, A. H. Mokdad, The phq-8 as a measure of current depression in the general population. J. Affect. Disord. **114**, 163–173 (2009)
46. K. L. Smarr, A. L. Keefer, Measures of depression and depressive symptoms: beck depression inventory-ii (bdi-ii), center for epidemiologic studies depression scale (ces-d), geriatric depression scale (gds), hospital anxiety and depression scale (hads), and patient health questionnaire-9 (phq-9). Arthritis Care Res. **63**(S11), 454–466 (2011)
47. S. Kung, R. D. Alarcon, M. D. Williams, K. A. Poppe, M. J. Moore, M. A. Frye, Comparing the beck depression inventory-ii (bdi-ii) and patient health questionnaire (phq-9) depression measures in an integrated mood disorders practice. J. Affect. Disord. **145**(3), 341–343 (2013)
48. K. Kroenke, R. L. Spitzer, J. B. Williams, The phq-9: validity of a brief depression severity measure. J. Gen. Intern. Med. **16**(9), 606–613 (2001)
49. F. Eyben, M. Wöllmer, B. Schuller, in *Proceedings of the 18th ACM International Conference on Multimedia*. Opensmile: the munich versatile and fast open-source audio feature extractor (Association for Computing Machinery, New York, 2010), pp. 1459–1462
50. C. Ding, H. Peng, Minimum redundancy feature selection from microarray gene expression data. J. Bioinforma. Comput. Biol. **3**(02), 185–205 (2005)
51. M. Valstar, B. Schuller, K. Smith, F. Eyben, B. Jiang, S. Bilakhia, S. Schnieder, R. Cowie, M. Pantic, in *Proceedings of the 3rd ACM International Workshop on Audio/visual Emotion Challenge*. Avec 2013: the continuous audio/visual emotion and depression recognition challenge (Association for Computing Machinery, New York, 2013), pp. 3–10
52. F. Ringeval, B. Schuller, M. Valstar, J. Gratch, R. Cowie, S. Scherer, S. Mozgai, N. Cummins, M. Schmitt, M. Pantic, in *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*. Avec 2017: real-life depression, and affect recognition workshop and challenge (Association for Computing Machinery, New York, 2017), pp. 3–9

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Submit your manuscript to a SpringerOpen journal and benefit from:

► Convenient online submission
► Rigorous peer review
► Open access: articles freely available online
► High visibility within the field
► Retaining the copyright to your article

Submit your next manuscript at ► springeropen.com