Audio-based Detection of Anxiety and Depression via Vocal Biomarkers

Raymond Brueckner, Namhee Kwon, Vinod Subramanian, Nate Blaylock, and Henry O'Connell

Canary Speech Inc., Provo UT 84604, USA
{ray, namhee, vinod, nate, henry}@canaryspeech.com,
https://www.canaryspeech.com

Abstract. We present a comparison of results based on the application of various model/feature combinations on the task of detecting anxiety and depression from audio signals of spontaneous speech. The adopted models comprise several different advanced deep neural networks, including CNN, LSTM, and attention networks, and are compared against traditional, shallow machine learning models. As input features we compare supra-segmental, paralinguistic feature sets against classical Mel-Frequency Cepstral Coefficients and advanced pre-trained X-vector and Wav2Vec2 features. Our models are trained based on self-assessment scores: GAD-7 for anxiety and PHQ-8 for depression. We present binary classification results for anxiety and depression separately and show that despite the noisy self-assessment labels our best model is able to achieve an unweighted average recall (UAR) of 0.60 for anxiety and 0.63 on the depression task. The result on the anxiety task almost reaches the reported self-scored GAD-7 screening reliability of 0.64. This shows that our best audio-based model can be deployed as an anxiety and depression screening tool.

Keywords: audio analysis, anxiety detection, depression detection, mental health, wav2vec2, x-vector

1 Introduction

According to estimates of the World Health Organization (WHO) roughly 280 million people of the world's population suffer from depression and around 301 million people from anxiety disorders[1]. Further, that source states that many people also do not have access to healthcare support due to cost and infrastructure barriers. As a solution to this situation the field of *remote sensing for healthcare* has recently been gaining popularity [9] as an interdisciplinary field of computer science and medicine. Leveraging mobile devices and IoT to record

[1] https://www.who.int/news-room/fact-sheets/detail/mental-disorders

data and assess individuals' health statuses, the aim is to render healthcare services more accessible to everyone.

In this context, we focus on conducting research and developing automated solutions for detecting, monitoring, and screening health conditions. These solutions can be leveraged in a multitude of scenarios, such as in clinical situations, which require early diagnosis, health check-ups, monitoring health development, or supporting long-term treatment. Automated solutions can not only support and detect cases that might otherwise remain undetected, but they can also save the human labor, time, and money of clinicians and institutions. Eventually, there is an increasing interest and need for wellness applications in order to detect potentially problematic health issues as early as possible.

This study focuses on mental healthcare, specifically on the detection of *anxiety* and *depression*. While prior work in this area adopted multi-modal approaches, i. e. audio and video features [43][31], in this study, we focus exclusively on the detection via audio-based vocal biomarkers. This allows us to obtain insights into how much there is to be gained from that single modality. Moreover, it provides a non-invasive, simple, and cost-effective method to collect and assess participants' health status in the above regard. We point out that we refer to anxiety as *trait* anxiety, which is a measure of the stable tendency of the anxiety in a person, separate from a momentary situation one might be in [10]. This is in contrast to other studies which analyzed *state* anxiety, e. g. anxiety during public speech [22][30]. Trait anxiety may manifest more pervasively in speech compared to momentary or state anxiety.

We will describe the conditions of the data set used in this study, including the audio recordings and their respective self-assessment labels, in Section 2, followed by a description of our general approach and used measures to report performance in Section 3. We will then describe the features derived and used in the experiments in Section 4, especially since we use partially less known or less used features in this study. We finally describe and discuss experiments and their results in Secion 5, followed by our conclusions in Section 6.

2 Data

The data used in this work were collected from different proprietary sources and a multitude of persons who were asked to respond to the question "How are you?". This is in contrast to other studies which used specially designed questions or prompts [19]. The main rationale for this choice is that we wanted to obtain natural, free speech, without any specific lexical or semantic constraints as would happen e. g. in read speech, so that this application can be adopted in broader use cases. Since in this study we deal with an audio-only based approach, this could be especially important, as free speech exhibits different characteristics regarding speaking style - such as prosody, pause frequency and lengths, rate of

speech, etc. - than read speech. Furthermore, one does not have to deal with any potential impact of the text to be read, which could introduce an undesirable bias due to the person's education level or reading proficiency. The participants in this study were recruited with the goal to obtain good coverage and balance in gender, age, race, state of residence, state of childhood, etc.

2.1 Audio Recordings

The audio data were collected in different audio environments, however, all with a reasonably high Signal-to-Noise Ratio (SNR) and in relatively "clean" situations, i. e. with no to only small background noise. The audio recordings were stored in uncompressed PCM/WAV format at a sampling rate of 16 kHz, typical in speech-based scenarios and sufficient to capture the frequencies of the human voice.

After data cleaning, we ended up with audio recordings from 4748 participants in total, accompanied by their respective self-reported answers to the anxiety and depression questionnaires (cf. Section 2.3). Since not all participants reported to both questionnaires the cardinality for the two domains differs. Table 1 shows the relevant statistics for the respective distributions.

Table 1: Statistics over the audio recordings for anxiety and depression

	Anxiety	**Depression**
Nr. audio recordings	4748	4405
Total duration	84h 17m 26s	79h 37m 16s
Max. duration	4m 32s	4m 32s
Mean duration (μ)	1m 4s	1m 5s
Stddev. duration (σ)	17 s	16 s

2.2 Voice Activity Detector

Since in this study we only focus on purely audio-based speech characteristics and do not investigate more speech-related aspects such as duration, frequency, or length of pauses and speech segments, fillers, repetitions, etc., we apply a voice activity detector (VAD), which is able to differentiate between speech and non-speech segments. This is furthermore needed since some of our recordings show overly long pause segments in front of, during, or at the end of the speech, and processing pauses could potentially skew our prediction outcomes and deteriorate system performance. Therefore, all descriptions and results presented in this study are based on audio signals after applying the VAD and filtering out the silence regions.

2.3 Annotation: Self-assessment GAD-7 and PHQ-8 Scores

In order to train supervised models and evaluate their performance, the acoustic data must obviously be annotated by the target value or class to be predicted by the model. In this context one needs to differentiate between a *ground truth*, i. e. the objectively verifiable class membership[2], and the *gold standard*, which is the result of the annotation process [32]. Ideally one wishes the former to be identical to the latter, in reality, this is not always the case. For example, it is (relatively) easy to annotate the gender and age of people, but the annotation of complex and difficult-to-observe phenomena like anxiety or depression, can lead to ambiguous labels, even if assessed by multiple clinicians. This undesired variability, often termed *label noise*, is often approached by annotating the data (in our case audio recordings) by multiple annotators and then deriving the *gold standard* by aggregation of the individual annotations, e. g. via majority voting. However, this approach is very costly and sometimes unfeasible.

For this reason, our study relies on the self-reported scores (*self-assessment)* by the participants as the gold standard label. We adopt the GAD-7 questionnaire for anxiety and PHQ-8 questionnaire for depression. The GAD-7 (General Anxiety Disorder-7) is a standard self-reported scale for screening anxiety disorder, which was proposed by Spitzer et al. in 2006 [41]. It is composed of 7 questions asking how often the person has been bothered by specific feelings over the past 2 weeks. The total score, y, ranges between 0 to 21, calculated by summing each question's answer which is on a 4-point scale (0-3). Spitzer suggested cutoff scores for 4 categories to interpret the GAD-7 scores: minimal ($y < 5$), mild ($y < 10$), moderate ($y < 15$), and severe ($y \geq 15$).

A study by Beard et al. in 2014 [5] compared the efficacy of the GAD-7 for anxiety disorder detection against a mental health professional's diagnosis and reported sensitivity and specificity of 0.74 and 0.54, respectively. *Sensitivity* and *Specificity* are commonly used in the medical field to evaluate diagnostic efficacy. They are defined as the proportion of correctly diagnosed positive and negative labels, respectively. In the current bio-medical context *positive* commonly refers to the target class of interest, e. g. *depressed* or *suffering from GAD*, while *negative* refers to the disjunct healthy, control group. Hence, sensitivity is the recall of positive labels (class label 1) and specificity is the recall of the negative labels (class label 0), where the recall of class c is defined as

$$Recall(c) = \frac{N_{TP}^{(c)}}{N_{TP}^{(c)} + N_{FN}^{(c)}} \tag{1}$$

which is the ratio between the number of correctly predicted instances of that class, $N_{TP}^{(c)}$, over the number of all instances from that class, $N_{TP}^{(c)} + N_{FN}^{(c)}$ [42]. Hence, sensitivity is $Recall(1)$ and specificity is $Recall(0)$. By definition, the

[2] One could, for example, assign a person to the *depressed* class based on independent depression screening tests.

value range for both is $Recall(c) \in [0, 1]$. Note that the comparison between the self-reported GAD-7 scores and the diagnoses of a mental health professional in [5] was done for the binary case i. e. generalized anxiety disorder (GAD) for $y \geq 10$ and healthy or minimal GAD for $y < 10$.

Similarly, the PHQ-8 is an 8-item questionnaire on depressive symptoms. It is based on the PHQ-9 (Patient Health Questionnaire-9) by excluding its final question, which asks about suicidal thoughts. The PHQ-9 [18] is a standard screening questionnaire, with answers on a 4-point scale for each item, resulting in a total sum in-between 0 and 27 (24 for PHQ-8). Previous research [48] showed a very strong correlation of 0.996 between PHQ-8 and PHQ-9. Both PHQ-8 and PHQ-9 suggest a cut-off score of 10 for screening major depressive disorder. PHQ-9 is reported to have a sensitivity of 0.74 and a specificity of 0.91 using the suggested cut-off score when compared to a health professional's diagnosis of depression disorder [2].

We report the details of the label distribution of our data set after applying the suggested cut-off score in Table 2. There are fewer examples with label categories *moderate* and *severe* in the data, for both GAD-7 and PHQ-8, which is consistent with the frequency of these disorders in the general population [4][47]. In this study, we investigate binary classification models which attempt to discriminate between the combination *Minimal*/*Mild* and the combination *Moderate*/(*Moderately Severe*)/*Severe*.

3 General Approach and Measures of Success

One prototypical approach to detect bio-medical, person-related states such as anxiety or depression via automated ML/AI systems is to adopt *sequence classification* of the underlying audio signal. This effectively means that a variable-length audio signal needs to be mapped to a single prediction output determining the class membership of that signal, for example, *depressed* or *non-depressed*. A big challenge in this approach is, however, that there is only very little information guiding the (supervised) learning process for a potentially very long input signal. To be more specific, we deal with binary classification - *depressed* vs. *non-depressed* or *GAD* vs. *no-GAD* - and hence there is one bit of information at the output of an ML network, which can be used to determine if the predicted output is equal to the "true" value (gold standard label). At the same time, the input sequences (cf. Section 4) often are composed of hundreds to thousands of feature vectors. The ML/AI network essentially needs to learn a mapping between these long, variable-length input sequences and that single bit of output target value. This generally poses severe challenges when training the networks of choice. One should note that this is in stark contrast to more commonly known domains like Automatic Speech Recognition (ASR) or Natural Language Processing (NLP) which have considerably more information to guide the learning process, e. g. the identity of words, phonemes, lexical elements, etc.

Table 2: Distribution of class labels

Class	**Anxiety (GAD-7)**	**Score**	**Count (%)**
Control	Minimal	$0 \leq y < 5$	1569 (33%)
	Mild	$5 \leq y < 10$	1573 (33%)
GAD	Moderate	$10 \leq y < 15$	1034 (22%)
	Severe	$15 \leq y \geq 21$	572 (11%)
Total			4748

(a) Anxiety (GAD-7)

Class	**Depression (PHQ-8)**	**Score**	**Count (%)**
Control	Minimal	$0 \leq y < 5$	1505 (34%)
	Mild	$5 \leq y < 10$	1296 (29%)
	Moderate	$10 \leq y < 15$	884 (20%)
Depressed	Moderately Severe	$15 \leq y < 20$	511 (12%)
	Severe	$20 \leq y \leq 24$	209 (5%)
Total			4405

(b) Depression (PHQ-8)

One possible countermeasure to alleviate the problem of limited information availability is to adopt pre-trained networks as part of the full network. Typically, these pre-trained networks have been trained on much larger data sets, often on data of a different domain or task, and hence capable of encoding the underlying audio signal into a robust and general intermediate representation for further use. Therefore, these pre-trained networks are commonly used in the early stages of a more complex system, followed by a so-called *downstream head* network. That latter network often can then be chosen to have fewer learnable parameters, which helps to combat overfitting and generalization issues. In Section 4 we will describe in more detail two typically performant variants of such pre-trained networks, namely X-Vector and Wav2Vec networks.

For the evaluation of performance *Accuracy* (Acc) is typically used in the ASR/NLP field. However, it suffers from a problem often encountered in imbalanced data sets, i. e. where one class contains (considerably) more examples than the other class(es), which is also the case in our study (cf. Table 2). The stronger the imbalance the more accuracy tends to reflect the performance of the majority class since it is a weighted accuracy. For this reason in areas where

relatively small data sets predominate, such as the bio-medical field or in the paralinguistics field in general, researchers usually prefer and report the *Unweighted Average Recall* (UAR), which is defined as

$$UAR = \frac{1}{C} \sum_{c=1}^{C} Recall(c) \tag{2}$$

where $Recall(c)$ is the recall specific of class c as defined in Eq. 1 and C is the number classes, $C = 2$ in the binary case. The UAR weights correct and incorrect predictions for all classes equally. Note that the UAR is sometimes also referred to as unweighted accuracy (UA) [33].

In the bio-medical field, one is not solely interested in the overall, averaged performance, but also in the specific performance in the detection of the target class, e. g. *depressed* or *GAD*, as well as the (healthy) control group. For this purpose, two specific measures are commonly adopted. Assuming that "positive" refers to the target class (with class label 1) and "negative" refers to the (healthy) control group (with class label 0), then: (a) *Sensitivity*, also termed true positive rate, is the ratio of positive predictions to the number of actual positives in the data. It is hence identical to the recall of the positive class, $Recall(1)$. (b) Likewise, *Specificity* is the ratio of negative predictions to the number of actual negatives in the data, and therefore identical to $Recall(0)$. The (un-weighted) average of Sensitivity and Specificity is equal to the UAR, but evaluating these two measures allows us to assess the model performance for the classes individually.

4 Features

4.1 Supra-segmental Features

As described in Section 2.1, the length of audio signals can be relatively large, typically on the order of 30-60 seconds or longer. In typical frame-based feature approaches (cf. following sections) this leads to thousands of feature vectors as an intermediate representation of the audio signal and not all model types are suitable to process that many frames. This fact is one of the reasons that in the area of computational paralinguistics[3] so-called *supra-segmental* features are also applied, which are one-dimensional vector representations of the full audio signal.

An illustration of the general feature extraction and prediction processing chain is given in Figure 1(b). In the first step frame-level features, so-called *low-level descriptors* (LLD), are extracted, typically over a short time window of duration 25 ms with a frame shift of 10 ms. Depending on the specific task-related

[3] Paralinguistics is the study of paralanguage, which connotes "alongside language" and generally describes the non-verbal elements of human communication, i. e. all meta-information that accompanies and complements language [6].

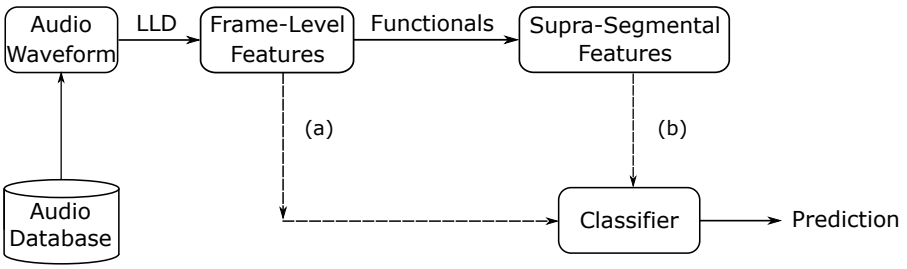

Fig. 1: Illustration of the general feature extraction and prediction processing chain.

needs the resulting features comprise (a) energy-related LLDs, e. g. loudness, RMS energy, zero-crossing rate, etc.; (b) spectral and cepstral LLD, such as Auditory Spectrum, Mel-Frequency Ceptral Coeffiecients (MFCC), Spectral Flux, Spectral Psychoacoustic Sharpness, etc.; (c) Prosodic/Voiding-related LLDs, including fundamental frequency (F_0), probability of voicing, jitter, shimmer, etc. One can then apply a number of *functionals* to a window of stacked LLDs, where the window can be chosen to include all LLD vectors into a single super-vector, the *supra-segmental* features. This approach also has the desirable effect of reducing the variable-length audio recording into a fixed-length representation, which is very helpful for models which are unable to handle time information, e. g. Support Vector Machines (SVM) or Random Forests (RF) (cf. subsequent sections below).

Examples of functionals are the minimum, maximum, mean, standard deviation, or higher-order moments such as skew and kurtosis. Further, percentiles, temporal centroids, regression coefficients, peaks and valleys are commonly used. Furthermore, different functionals are applied to different features or feature groups. For a detailed description and further information the interested reader is referred to [33][11][6].

In the recent decade supra-segmental features have often been used with shallow machine learning models as a baseline reference for more advanced and complex ML models, for example in numerous Interspeech Paralinguistic Challenges [34][35][38][37]. For that same reason we experimented with a few variants of that feature type, which are described in this section. All supra-segmental features were extracted with the openSMILE toolkit [13] and we chose feature sets of considerably different dimensionality to assess whether the different features and the information contained therein has an impact on classification performance.

4.1.1 eGeMAPS

The *extended Geneva Minimalistic Acoustic Parameter Set* (eGeMAPS) was first proposed in [12] and consists of 88 features per audio recording. The extracted LLDs are smoothed over time with a symmetric moving average filter of length 3, and the arithmetic mean and coeffcient of variation (standard deviation normalized by the arithmetic mean) are applied as functionals. Additional functionals are applied to loudness and pitch. Furthermore, the average RMS energy and 6 temporal features are included, which are: the rate of loudness peaks per second, mean length and standard deviation of continuous voiced and unvoiced segments, and the rate of voiced segments per second, which approximates the pseudo-syllable rate [29]. Note that some functionals are applied only to either voiced or unvoiced regions, depending on the type of the LLD. The interested reader is referred to [11] for the exact details on how the features are computed.

4.1.2 AVEC 2013

The AVEC 2013 feature set was developed for the acoustic part of *The Continuous Audio/Visual Emotion and Depression Recognition Challenge* and consists of 2,268 features composed of 32 energy and spectral-related LLDs x 42 functionals, 6 voicing-related LLDs x 32 functionals, 32 delta coefficients of the energy/spectral LLDs x 19 functionals, 6 delta coefficients of the voicing related LLDs x 19 functionals, and 10 voiced/unvoiced durational features [44]. Contrary to the sliding window approach over 20 second-long segments suggested in [44], we compute one supra-segmental feature vector on the full audio signal to be able to use it with shallow models without any further post-processing. Note the considerably increased dimensionality of the AVEC 2013 feature set w. r. t. the eGeMAPS feature set from Section 4.1.1.

4.1.3 ComParE 2016

Among the three supra-segmental feature sets adopted in this study the *Interspeech 2016 COMputational PARalinguistics challengE* (ComParE 2016) [38] feature set is the largest, containing 6,373 features altogether and almost three times larger than the AVEC 2013 feature set. It is the result of many years of research and its development and availability has been fundamental for the objective evaluation and comparison of many different algorithms in the area of paralinguistic research [36]. Due to the complexity of its derivation, we only briefly describe its components: based on 65 LLDs (4 energy-related, 55 spectral/cepstral, and 6 voicing-related), a large range of functionals are applied to a sequence of LLD vectors, including statistical moments, percentiles, extrema, peaks/valleys, up-level times, rise/curvature times, segment lengths, regression and linear prediction coefficients, etc. Again, we invite interested readers to consult the more detailed explanations in [13][6].

4.2 Frame-/Segment-level features

Contrary to the supra-segmental features described above, frame- or segment-level features effectively are identical to the underlying LLDs, i. e. they are computed over short, shifted segments of the audio signal. They are directly fed into any downstream model without any further postprocessing, as depicted in Figure 1(b).

4.2.1 Mel-Frequency Cepstral Coefficients

Among the multitude of possible frame-level features, we limit our investigation to Mel-Frequency Cepstral Coefficients (MFCC), which were the default features used in the field of ASR for decades and have also seen widespread use in the fields of speaker recognition, music information retrieval, and others. MFCCs are essentially derived by computing the Fourier transform on a windowed segment of audio input, mapping the power spectrum onto the Mel scale, filtering the resulting representation by triangular (or similarly shaped) filters, applying the logarithm of each filter output, and computing the discrete cosine transform (DCT) on the output. The MFCCs are the magnitudes of the resulting cepstrum [8].

The DCT approximately decorrelates the MFCCs and by choosing the number of MFCCs to be smaller than the number of critical Mel bands, one obtains some sort of feature reduction. This is common since most of the information is contained in the lower-order coefficients. For example, MFCC(0), the first MFCC coefficient, which accounts for the distribution of high vs. low-frequency components in the Mel-frequency spectrum, is highly informative for the prediction of arousal in a number of audio applications [46].

4.2.2 X-vector Features

X-vectors were initially proposed for the task of speaker identification [40]. These vectors encode several aspects of a speaker's voice [28] and have also been used in other application domains, such as language identification [39], sentiment evaluation [24][25], but also in the bio-informatics field for the detection of early symptoms of Parkinson's and Alzheimer's disease [16][20][23], and the detection of depression [author].

In this study, we follow a similar approach as in [author] We use the publicly available 16 kHz English VoxCeleb x-vector model[4] [27], pre-trained on over 1 million audio recordings from more than 7000 celebrities from the VoxCeleb 1 and 2 data sets. The input to the model are 30-dimensional Mel Frequency Cepstral Coefficients (MFCC), extracted from the audio recordings (sampling rate f_s = 16000 Hz) using a frame length of 25 ms and a hop size of 10 ms. The MFCC

[4] https://kaldi-asr.org/models/8/0008_sitw_v2_1a.tar.gz

frames were normalized using Cepstral Mean Variance Normalization (CMVN) before being fed to the network.

The x-vector model itself is a deep neural network consisting of five 1-dimensional time-delay neural networks (TDNN) layers [45], a statistics pooling layer, and one (depression) or two (anxiety) fully connected (FC) layers. We chose the number of output layers depending on the best performance of the full network. All convolutional TDNNs used the rectified linear unit (ReLU) activation function [21] and batch normalization [15]. This network generates a variable-length sequence of feature vectors of dimensionality 3000. The subsequent statistics pooling layer computes the mean and standard deviation over segments of 150 frames, equivalent to 1.5 seconds of audio. The resulting fixed-length representation can now be fed into a fully-connected feed-forward neural network to generate 512-dimensional output vectors, which is termed *x-vector*. Informal experiments have shown that averaging four consecutive x-vectors leads to a more robust representation, hence the output of the described network covers a minimal audio length of 6 seconds. The entire model has approximately 4.2 million parameters, which are frozen and not further fine-tuned during training (cf. Section 5).

4.2.3 Wav2Vec2 Features

Wav2Vec2 was originally proposed as a framework for self-supervised learning of meaningful speech representations on large amounts of unlabeled audio data, with applications to ASR and NLU [3]. Contrary to traditional ASR approaches that operate on hand-engineered features, such as spectrograms or MFCCs (cf. Section 4.2.1), Wav2Vec2 directly processes the raw audio signal. By employing a contrastive learning framework, Wav2Vec2 learns to predict future audio samples given a set of preceding samples, thereby capturing intricate patterns and contextual information present in the audio signals. These learned representations can then be extracted and utilized as intermediate features for downstream speech processing tasks, in our case anxiety and depression detection. Wav2Vec2 extracts audio segments via a sliding window approach, which are then encoded by a multi-layer convolutional neural network (CNN) to create a latent representation. This representation is subsequently fed to a Transformer network to build contextualized representations.

The integration of Wav2Vec2 as intermediate feature representations offers several advantages. Firstly, it eliminates the need for hand-engineered features, as the model learns directly from the raw audio data; this has been reported to lead to improved generalization and adaptability across various speech processing domains. Secondly, it can effectively capture fine-grained acoustic details and linguistic content. Furthermore, Wav2Vec2 exhibits transfer learning capabilities, allowing it to leverage the model's pre-trained weights and fine-tune it on specific speech processing tasks with limited labeled data. This transfer learning

approach not only reduces the data requirements but also helps in mitigating the domain mismatch problem often encountered in real-world scenarios.

In our study, we used the publicly available pre-trained Wav2Vec2 model *facebook/wav2vec2-large-robust* from Huggingface.co, which generates a 512-dimensional feature vector approximately every 20 ms.

5 Experiments and Results

In the following we present our current results on anxiety and depression detection based solely on audio input. As already mentioned in the introduction, we do not take into account more speech-specific features such as pause/silence, voice/unvoiced, lexical, or other high-level information. Instead, we only rely on the feature sets described above and what they represent about the input audio.

The results we present are based on the best models we found after extensive hyper-parameter tuning. In order to guarantee meaningful results we randomly split our data set into training, development, and test sets with a ratio of approximately 8:1:1, attempting an equal label distribution across all splits. Since there is only one audio recording per participant in our data set it is guaranteed that a test participant is never seen during training or during hyper-parameter tuning. The size of the resulting blind test is 406 for anxiety and 405 for depression and all results presented are derived from the test set.

5.1 Shallow Models and Supra-Segmental Features

Shallow machine learning models refer to a class of machine learning algorithms that typically have a limited number of layers or a simple structure. While shallow models may not achieve the same level of performance as deep learning models on highly complex tasks, they have the advantage of generally being computationally efficient, and easier to interpret. We decided to include experiments with them in combination with the supra-segmental features described in Section 4.1 for two reasons: first, to serve as a baseline performance reference; and more importantly second, because given the limited amount of our data examples (just one feature vector per audio recording), deep models could exhibit overfitting.

We compare three types of shallow models, namely *Support Vector Machines* (SVM), *Random Forests* (RF), and *XGBoost*, which are all known to be robust against overfitting and offer good generalization performance in many domains. SVM and Random Forest models were trained with the SKLearn toolkit [26] and XGBoost with the Python xgboost package [7]. For XGBoost we also tried two variants, namely the *linear* and the *tree* versions. Since all models come with a large variety of hyper-parameters we performed an extensive tuning using the SKLearn hyper-parameter optimizer over all relevant parameters per

Table 3: Test set results of the shallow models and supra-segmental features

Model	**Feature Set**	**Anxiety**				**Depression**			
		Acc	**UAR**	**Sens**	**Spec**	**Acc**	**UAR**	**Sens**	**Spec**
SVM	eGeMAPS	0.53	**0.56**	0.62	0.49	0.59	0.58	0.54	0.61
SVM	AVEC	0.54	0.54	0.53	0.54	0.60	0.59	0.56	0.61
SVM	ComParE	0.57	**0.56**	0.52	0.59	0.61	**0.60**	0.56	0.64
Random Forest	eGeMAPS	0.65	0.52	0.11	0.93	0.64	0.54	0.20	0.89
Random Forest	AVEC	0.66	0.52	0.10	0.95	0.63	0.53	0.15	0.91
Random Forest	ComParE	0.64	0.51	0.09	0.93	0.64	0.55	0.18	0.91
XGBoost (Linear)	eGeMAPS	0.45	0.51	0.68	0.34	0.56	0.57	0.61	0.53
XGBoost (Linear)	AVEC	0.52	0.54	0.61	0.47	0.62	**0.59**	0.48	0.70
XGBoost (Linear)	ComParE	0.57	**0.56**	0.54	0.59	0.58	0.58	0.59	0.58
XGBoost (Tree)	eGeMAPS	0.61	0.51	0.20	0.82	0.61	0.56	0.37	0.75
XGBoost (Tree)	AVEC	0.66	0.55	0.22	0.88	0.64	**0.59**	0.40	0.77
XGBoost (Tree)	ComParE	0.61	0.54	0.34	0.74	0.59	0.53	0.29	0.77

model type. As we used 5-fold cross-validation in this optimization process, we merged the train and validation data splits.

Table 3 shows the test set results for the best models after hyper-parameter tuning for each supra-segmental feature set. From the UAR results, it is difficult to conclude which feature set proves to be optimal. There is a tendency that XGBoost performs slightly better than Random Forests, esp. on the depression task. However, SVMs perform better or at last equally well on both the anxiety and the depression task for almost all feature sets and the ComParE 2016 feature set - which has the most features and hence presumably contains the most information about the underlying audio signal - seems a good choice for this shallow model type. Another advantage of SVM models is the relative balance between Sensitivity and Specificity, which is a desirable characteristic esp. in bio-medical and clinical scenarios.

The results for both tasks, $UAR = 0.56$ for anxiety and $UAR = 0.60$ for depression, are significantly above the chance level ($UAR = 0.50$), yet compared to other tasks in the literature they still seem relatively low. Nonetheless, our results compare favorably to published results on other academic data sets. For example, Huang et al. [14] reported accuracies of $Acc = 0.59$ on DAIC-WoZ and $Acc = 0.63$ on FORBOW[5]. While we suggest UAR as a target measure (as

[5] Note that this data set was labeled by trained clinical assessors, not relying on self-assessment labels.

explained in Section 3, it gives an approximate indication that the results on all data sets might be approximately comparable. This fact suggests that the purely audio-based detection of anxiety and depression is very challenging.

5.2 Deep Models and Frame-Level Features

In this section, we describe the results of adopting advanced, state-of-the-art, deep neural network models for the problem of detecting anxiety and depression from audio recordings. These models are not only generally more powerful because they allow modeling more intricate and complex patterns in the underlying audio, but also they allow to use sequential, frame- or segment-level input data. However, due to their complexity and higher number of trainable parameters, compared to the shallow models above, they run the risk of overfitting more easily, which must be combated, e. g. by adopting regularization or normalization techniques during training and topology definition.

In this study, we present a small selection of the best-performing feature/model combinations on our current data set, which we briefly describe here:

- Pooling + Feed-Forward Deep Neural Network (FF-DNN): This approach is the simplest of the presented performing a mean pooling of all feature vectors per audio input and feeding the resulting vector into a two-layer FF-DNN, each layer of size 256 with ReLU activation function and followed by Layer Normalization.

- Long Short-Term Memory (LSTM) is an advanced recurrent neural network (RNN) proven to be highly effective even on longer time sequences, hence mitigating the vanishing gradient problem often encountered with regular RNNs. We trained the LSTM with 128 hidden states and collected the outputs of the LSTM at each time step and applied mean pooling to generate a fixed-length representation for the full audio input. We also tried bi-directional BLSTMs, but they did not show any improvement over the uni-directional LSTMs.

- Attention + LSTM: This adds an attention layer in front of the LSTM network described above. The idea behind it is that through the attention layer the network should learn on which parts of the input to put its focus. Relevant parts should get a higher weight w. r. t. to other parts of the input. We used scaled, dot-product attention, 64-dimensional embedding matrices for *query*, *key*, and *value*. The attention outputs were fed into the LSTM with 64 hidden states and the outputs of each time step averaged over time via a 1-D global average pooling.

- Convolutional Neural Network (CNN) + FF-DNN: In this model configuration we used the (averaged) outputs of the (fixed) X-Vector model described in Section 4.2.2, and fed them into 2-lacer stacked CNN network, where each

layer has 256 filters, a kernel size of 3, and a ReLU activation function. We also used L_2 regularization ($p = 0.01$), dropout, and batch normalization during training. The outputs of this network are averaged along the time axis via generalized mean pooling ($p = 3$) to yield a fixed-size representation for the audio segment under consideration. That output vector is eventually fed into a dense layer of size 128 with ReLU activation. Since the composed model outputs a sequence of predictions, we compute the final prediction by majority voting of the individual prediction results in the sequence.

Note that each output vector of the above models is fed into a single-layer neural network with a softmax output for class discrimination.

All models were trained on the train split using the TensorFlow framework [1]. We used the Adam [17] optimizer and the categorical cross-entropy loss as the objective function. Learning rates were set to $10 - e3$ after preliminary experiments. The input data was processed in randomized batches of size 10. Early stopping, hyper-parameter tuning, and model selection were performed on the validation split.

Table 4: Test set results of the deep models and segment-level features

Model	**Feature Set**	**Anxiety**				**Depression**			
		Acc	**UAR**	**Sens**	**Spec**	**Acc**	**UAR**	**Sens**	**Spec**
(a) LSTM+Pool	MFCC	0.59	0.53	0.35	0.72	0.60	0.59	0.52	0.65
(b) LSTM+Pool	Wav2Vec2	0.67	**0.60**	0.38	0.82	0.63	**0.63**	**0.63**	0.63
(c) Att+LSTM+Pool	Wav2Vec2	0.63	0.52	0.18	0.0.86	0.64	0.59	0.38	0.79
(d) Pool+FF-DNN	Wav2Vec2	0.62	0.54	0.29	0.79	0.64	**0.62**	0.53	0.71
(e) CNN+FF-DNN	X-Vector	0.66	0.58	0.42	0.74	0.67	**0.62**	0.49	0.75

Table 4 shows the test set results for the best models, from which a number of interesting conclusions can be drawn. First, comparing (a) and (b) one can see that the Wav2Vec2 features are clearly superior to the MFCC features, resulting in a significant gain on both tasks. On the depression task, it also shows a perfect balance between Sensitivity and Specificity. Interestingly, the addition of an attention layer (d) unexpectedly does not bring any improvement to the LSTM model; on the contrary, it deteriorates UAR performance considerably. The reason for this is yet unclear and requires further investigation. One potential reason could be the large number of feature vectors the Attention module has to cope with. In the future, we will investigate a shorter, segment-based approach to evaluate if this is helpful.

Equally interesting is the fact that the conceptually simple feed-forward network (d) (in combination with a preceding pooling) achieves a UAR almost on par with model (b) on the depression task. However, there is a significant regression on the anxiety task of 6% absolute. Eventually, the CNN-based model (e) using x-vectors as the input only performs slightly worse than the best model (b), again with a UAR almost on par in the depression task and slightly worse on the anxiety task.

To summarize, the best model (b) in this study achieves a UAR of 0.60 on the anxiety task and 0.63 on the depression task, which are both higher than the chance level ($UAR = 0.5$). On the anxiety task, the reported self-scored GAD-7 screening reliability [5] is $UAR = 0.64$, and hence our results are very close. However, as reported in a previous study by [author] the performance of the depression model of 0.63 is still considerably worse than the PHQ-9 self-label screening performance of 0.83 [2].

6 Conclusions

In this study we presented a comparison of various shallow and deep models in combination with supra-segmental and segment-level features for the audio-based detection of (trait) anxiety and depression from vocal biomarkers. We showed that deeper, more complex neural network models combined with pre-trained intermediate audio representation, such as Wav2Vec2 or X-Vector features, outperform their shallow counterparts by a fair margin. Our best models obtain an unweighted average recall (UAR) of 0.60 for anxiety and 0.63 for the depression task. The result on the anxiety task falls short of the reported self-scored GAD-7 screening reliability of 0.64 just by a small margin and hence shows that this audio-based model can be deployed as an anxiety and depression screening tool.

Considering that our models are trained and evaluated on the self-measured, subjective, and hence potentially "noisy" labels, the model performance is highly meaningful and promising towards the goal of automatically and objectively identifying anxiety and depression disorders based on everyday speech, without the time-consuming task of answering the lengthy self-evaluating questionnaires.

We will continue to investigate more advanced models and features, in order to improve the detection of anxiety and depression. As part of that work, we plan to validate the results based on our data set against other databases in order to further evaluate the effect of self-assessment labels.

References

1. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., Zheng, X.: TensorFlow: Large-scale machine learning on heterogeneous systems (2015). URL https://www.tensorflow.org/. Software available from tensorflow.org
2. Arroll, B., Goodyear-Smith, F., Crengle, S., Gunn, J., Kerse, N., Fishman, T., Falloon, K., Hatcher, S.: Validation of PHQ-2 and PHQ-9 to screen for major depression in the primary care population. The Annals of Family Medicine **8**(4), 348 (2010). DOI 10.1370/afm.1139. URL http://www.annfammed.org/content/8/4/348.abstract
3. Baevski, A., Zhou, H., Mohamed, A., Auli, M.: Wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In: Advances in Neural Information Processing Systems 33 (NeurIPS 2020). Curran Associates Inc., Red Hook, NY, USA (2020)
4. Bandelow, B., Michaelis, S.: Epidemiology of anxiety disorders in the 21st century. Dialogues in Clinical Neuroscience **17**, 327–335 (2015)
5. Beard, C., Björgvinsson, T.: Beyond generalized anxiety disorder: Psychometric properties of the GAD-7 in a heterogeneous psychiatric sample. Journal of Anxiety Disorders **28**(6), 547–552 (2014). DOI https://doi.org/10.1016/j.janxdis.2014.06.002
6. Brueckner, R.: Application of deep learning methods in computational paralinguistics. Ph.D. thesis, Technische Universität München (2020)
7. Chen, T., Guestrin, C.: XGBoost: A Scalable Tree Boosting System. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16, pp. 785–794. ACM, New York, NY, USA (2016)
8. Davis, S., Mermelstein, P.: Comparison of Parametric Representations for Monosyllabic Word Recognition in Continuously Spoken Sentences. IEEE Transactions on Acoustics, Speech, and Signal Processing **28**(4), 357–366 (1980)
9. De Angel, V., Lewis, S., White, K., Oetzmann, C., Leightley, D., Oprea, E., Lavelle, G., Matcham, F., Pace, A., Mohr, D.C., et al.: Digital health tools for the passive monitoring of depression: a systematic review of methods. NPJ digital medicine **5**(1), 3 (2022)
10. Endler, N.S., Kocovski, N.L.: State and trait anxiety revisited. Journal of anxiety disorders **15**(3), 231–245 (2001)
11. Eyben, F.: Real-time Speech and Music Classification by Large Audio Feature Space Extraction. Springer (2015)
12. Eyben, F., Scherer, K.R., Schuller, B.W., Sundberg, J., André, E., Busso, C., Devillers, L.Y., Epps, J., Laukka, P., Narayanan, S.S., Truong, K.P.: The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing. IEEE Transactions on Affective Computing **7**(2), 190–202 (2016)
13. Eyben, F., Wöllmer, M., Schuller, B.: openSMILE – The Munich Versatile and Fast Open-Source Audio Feature Extractor. In: Proceedings of the 18th ACM International Conference on Multimedia, pp. 1459–1462. ACM, Florence, Italy (2010)
14. Huang, Z., Epps, J., Joachim, D.: Investigation of Speech Landmark Patterns for Depression Detection. IEEE Transactions on Affective Computing **13**(2), 666–679 (2022). DOI 10.1109/TAFFC.2019.2944380

15. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: International conference on machine learning, pp. 448–456. PMLR (2015)
16. Jeancolas, L., Petrovska-Delacrétaz, D., Mangone, G., Benkelfat, B.E., Corvol, J.C., Vidailhet, M., Lehéricy, S., Benali, H.: X-vectors: New quantitative biomarkers for early parkinson's disease detection from speech. Frontiers in Neuroinformatics **15**, 578,369 (2021)
17. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
18. Kroenke, K., Spitzer, R.L., Williams, J.B.W.: The phq-9: validity of a brief depression severity measure. Journal of General Internal Medicine **16**(9), 606–613 (2001). DOI 10.1046/j.1525-1497.2001.016009606.x. URL https://doi.org/10.1046/j.1525-1497.2001.016009606.x
19. Ma, X., Yang, H., Chen, Q., Huang, D., Wang, Y.: Depaudionet: An efficient deep model for audio based depression classification. In: Proceedings of the 6th international workshop on audio/visual emotion challenge, pp. 35–42 (2016)
20. Moro-Velazquez, L., Villalba, J., Dehak, N.: Using x-vectors to automatically detect parkinson's disease from speech. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1155–1159. IEEE (2020)
21. Nair, V., Hinton, G.E.: Rectified Linear Units Improve Restricted Boltzmann Machines. In: International Conference on Machine Learning (ICML) 2010, pp. 807–814 (2010)
22. Nirjhar, E.H., Behzadan, A., Chaspari, T.: Exploring bio-behavioral signal trajectories of state anxiety during public speaking. In: ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1294–1298. IEEE (2020). DOI 10.1109/ICASSP40776.2020.9054160
23. Pappagari, R., Cho, J., Moro-Velazquez, L., Dehak, N.: Using state of the art speaker recognition and natural language processing technologies to detect alzheimer's disease and assess its severity. In: INTERSPEECH, pp. 2177–2181 (2020)
24. Pappagari, R., Wang, T., Villalba, J., Chen, N., Dehak, N.: x-vectors meet emotions: A study on dependencies between emotion and speaker recognition. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 7169–7173. IEEE (2020)
25. Parra-Gallego, L.F., Orozco-Arroyave, J.R.: Classification of emotions and evaluation of customer satisfaction from speech in real world acoustic environments. Digital Signal Processing **120**, 103,286 (2022)
26. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)
27. Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., et al.: The kaldi speech recognition toolkit. In: IEEE 2011 workshop on automatic speech recognition and understanding, CONF. IEEE Signal Processing Society (2011)
28. Raj, D., Snyder, D., Povey, D., Khudanpur, S.: Probing the information encoded in x-vectors. In: 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), pp. 726–733. IEEE (2019)

29. Ringeval, F., Schuller, B., Valstar, M., Jaiswal, S., Marchi, E., Lalanne, D., Cowie, R., Pantic, M.: AV$^+$EC 2015: The First Affect Recognition Challenge Bridging Across Audio, Video, and Physiological Data. In: Proceedings of the 5th International Workshop on Audio/Visual Emotion Challenge, pp. 3–8. ACM, Brisbane, Australia (2015)
30. Sakib, M.N., Nirjhar, E.H., Feng, K., Behzadan, A., Chaspari, T., Chaspari, T.: Exploring individual differences of public speaking anxiety in real-life and virtual presentations. IEEE Transactions on Affective Computing pp. 1–1 (2021). DOI 10.1109/TAFFC.2020.3048299
31. Salekin, A., Eberle, J.W., Glenn, J.J., Teachman, B.A., Stankovic, J.A.: A weakly supervised learning framework for detecting social anxiety and depression. Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies **2**(2), 1–26 (2018)
32. Schuller, B.: Intelligent Audio Analysis – Speech, Music, and Sound Recognition in Real-Life Conditions. Habilitation thesis, Technische Universität München, Munich, Germany (2012)
33. Schuller, B., Batliner, A.: Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing. John Wiley & Sons, Chichester, UK (2014)
34. Schuller, B., Steidl, S., Batliner, A.: The INTERSPEECH 2009 Emotion Challenge. In: Proceedings of the 10th Annual Conference of the International Speech Communication Association (INTERSPEECH). ISCA, Brighton, UK (2009)
35. Schuller, B., Steidl, S., Batliner, A., Epps, J., Eyben, F., Ringeval, F., Marchi, E., Zhang, Y.: The INTERSPEECH 2014 Computational Paralinguistics Challenge: Cognitive & Physical Load. In: Proceedings of the 15th Annual Conference of the International Speech Communication Association (INTERSPEECH). Singapore (2014)
36. Schuller, B., Weninger, F., Zhang, Y., Ringeval, F., Batliner, A., Steidl, S., Eyben, F., Marchi, E., Vinciarelli, A., Scherer, K., Chetouani, M., Mortillaro, M.: Affective and Behavioural Computing: Lessons Learnt from the First Computational Paralinguistics Challenge. Computer Speech & Language **53**, 156–180 (2019). DOI 10.1016/j.csl.2018.02.004
37. Schuller, B.W., Batliner, A., Bergler, C., Mascolo, C., Han, J., Lefter, I., Kaya, H., Amiriparian, S., Baird, A., Stappen, L., Ottl, S., Gerczuk, M., Tzirakis, P., Brown, C., Chauhan, J., Grammenos, A., Hasthanasombat, A., Spathis, D., Xia, T., Cicuta, P., Rothkrantz, L.J., Zwerts, J.A., Treep, J., Kaandorp, C.S.: The INTERSPEECH 2021 Computational Paralinguistics Challenge: COVID-19 Cough, COVID-19 Speech, Escalation & Primates. In: Proceedings of the 22nd Annual Conference of the International Speech Communication Association (INTERSPEECH), pp. 431–435 (2021). DOI 10.21437/Interspeech.2021-19
38. Schuller, B.W., Steidl, S., Batliner, A., Hirschberg, J., Burgoon, J.K., Baird, A., Elkins, A.C., Zhang, Y., Coutinho, E., Evanini, K.: The INTERSPEECH 2016 Computational Paralinguistics Challenge: Deception, Sincerity & Native Language. In: Proceedings of the 17th Annual Conference of the International Speech Communication Association (INTERSPEECH), vol. 2016, pp. 2001–2005. ISCA, San Francisco, CA, USA (2016)
39. Snyder, D., Garcia-Romero, D., McCree, A., Sell, G., Povey, D., Khudanpur, S.: Spoken language recognition using x-vectors. In: Odyssey, pp. 105–111 (2018)
40. Snyder, D., Garcia-Romero, D., Sell, G., Povey, D., Khudanpur, S.: X-vectors: Robust dnn embeddings for speaker recognition. In: 2018 IEEE international con-

ference on acoustics, speech and signal processing (ICASSP), pp. 5329–5333. IEEE (2018). DOI 10.1109/ICASSP.2018.8461375
41. Spitzer, R.L., Kroenke, K., Williams, J.B., Löwe, B.: A Brief Measure for Assessing Generalized Anxiety Disorder: The GAD-7. Arch Intern Med. **166**(10), 1092–1097 (2006)
42. Ting, K.M.: Precision and Recall, pp. 781–781. Springer, Boston, MA, USA (2010)
43. Valstar, M.F., Gratch, J., Schuller, B.W., Ringeval, F., Cowie, R., Pantic, M. (eds.): Proceedings of the 6th International Workshop on Audio/Visual Emotion Challenge, AVEC@MM 2016. ACM, Amsterdam, The Netherlands (2016)
44. Valstar, M.F., Schuller, B.W., 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: B.W. Schuller, M.F. Valstar, R. Cowie, J. Krajewski, M. Pantic (eds.) Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge, AVEC@ACM Multimedia 2013, Barcelona, Spain, October 21, 2013, pp. 3–10. ACM (2013)
45. Waibel, A.H., Hanazawa, T., Hinton, G.E., Shikano, K., Lang, K.J.: Phoneme recognition using time-delay neural networks. IEEE Transactions on Acoustics, Speech, and Signal Processing **37**, 328–339 (1989)
46. Weninger, F., Eyben, F., Schuller, B.W., Mortillaro, M., Scherer, K.R.: On the Acoustics of Emotion in Audio: What Speech, Music, and Sound Have in Common. Frontiers in Psychology **4** (2013)
47. Werneck, A.O., Silva, D.R.: Population density, depressive symptoms, and suicidal thoughts. Revista Brasileira de Psiquiatria (2020). DOI 10.1590/1516-4446-2019-0541
48. Wu, Y., Levis, B., Riehm, K.E., et al.: Equivalency of the diagnostic accuracy of the PHQ-8 and PHQ-9: a systematic review and individual participant data meta-analysis. Psychological Medicine **50**(8), 1368–1380 (2020). DOI 10.1017/S0033291719001314