ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) | 979-8-3503-4485-1/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICASSP48485.2024.10446567

A MACHINE-LEARNING MODEL FOR DETECTING DEPRESSION, ANXIETY, AND STRESS FROM SPEECH

Mashrura Tasnim, Ramon Diaz Ramos, Eleni Stroulia

Department of Computing Science
University of Alberta, Edmonton, Canada

Luis A. Trejo

School of Engineering and Sciences,
Tecnologico de Monterrey, Atizapán, Mexico

ABSTRACT

Predicting mental health conditions from speech has been widely explored in recent years. Most studies rely on a single sample from each subject to detect indicators of a particular disorder. These studies ignore two important facts: certain mental disorders tend to co-exist, and their severity tends to vary over time. This work introduces a longitudinal dataset labeled with depression, anxiety, and stress scores using the DASS-21 self-report questionnaire, and describes a machine-learning pipeline to determine the severity of the three mental disorders using acoustic features extracted from speech samples of this dataset. Our initial findings suggest that healthy participants adhere more to the study procedure than participants who exhibit indicators of depression, anxiety, and stress and demonstrate that a one-dimensional convolutional neural network, trained on VGG-19 features, predicts the severity of depression, anxiety, and stress with high accuracy.

Index Terms— Depression, anxiety, stress, speech analysis, convolutional neural network.

1. INTRODUCTION

Mental disorders like depression, anxiety, and stress involve significant disturbances in thinking, emotional regulation, and behavior, affecting the individual's day-to-day life and well-being. 12.5% of the world population lives with a mental disorder, among which depression and anxiety are most common [1]. In recent years, speech has been considered a reliable biosignal for measuring the severity of high-priority mental disorders, including depression, anxiety, schizophrenia, stress, and Alzheimer's Dementia [1, 2, 3], because of its non-invasive nature and cost-effectiveness. Numerous machine learning models have been proposed by researchers to detect indicators of mental disorders [4, 5, 6, 7] etc. Although many of the mental disorders are interrelated and tend to co-exist, few studies propose prediction methodologies for co-morbid conditions, such as depression, anxiety, and stress. The lack of datasets labeled with scores of multiple conditions is a major cause of the scarcity. The Distress Analysis Interview Corpus (DIAC-WoZ) is a well-known speech corpus labeled with depression and post-traumatic stress disorder (PTSD) [8]. Different subsets of this dataset were introduced as the challenge corpus of the Audio-visual Emotion Recognition Challenge (AVEC) in 2016, 2017, and 2019 [9, 10, 11]. The DEPression and Anxiety Crowdsourced Corpus (DEPAC) is a recently published dataset with depression and anxiety labels on Patient Health Questionnaire–8 (PHQ-8) and General Anxiety Disorder-7 (GAD-7) scales, respectively [12]. Both of these datasets contain samples in the English language. To the best of our knowledge, there is no speech corpus labeled with depression, anxiety, and stress scores.

In this paper, we introduce a new longitudinal speech corpus containing over 1,000 speech samples in English and Spanish, collected from May 2022 to March 2023. COVID-19 pandemic has increased the prevalence of mental health disorders, and studies show that during the pandemic, youth (15 to 39 years) were more vulnerable to depression and anxiety disorders [13]. Our dataset was collected during the post-pandemic period, recruiting participants between 19 and 29 years old, which will be a valuable resource for the researchers in this area. In our dataset, we have observed a significant positive correlation among the disorder scores, which supports the fact that these disorders tend to be comorbid. Our analysis also shows a positive correlation between disorder severity and participants' adherence to study protocol, indicating that healthy individuals demonstrate more conformity to routine activities than persons with mental disorders. Finally, we formulate a Convolutional Neural network (CNN) for predicting depression, anxiety, and stress scores on Depression Anxiety Stress Scales (DASS-21) with root-mean-square errors of 7.09, 7.69, and 8.40 out of 42. The model's performance is competitive with the state-of-the-art.

2. THE DATASET

We collected speech samples from 40 participants between the ages of 19 to 29 years old from Mexico and Canada. 26 of the participants were native Spanish speakers (14 female, 9 male, and 3 identified as other gender), and 14 were English speakers (6 male, 8 female). Every three days, the participants

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

provided two speech samples: guided reading and free-form speech. For guided reading, the participant read out the paragraph 'Please call Stella' [14] in English or Spanish. For the free-form speech task, the participants were asked two questions, randomly selected from a list of questions, including describing a memorable event, a hobby, or a favorite person. We developed an Android application for data collection. The application prompted the participants to record samples every three days. The data collection continued for two months, resulting in a corpus of 1,049 data points. 838 of the samples are in Spanish, and the rest are in English. Each audio sample ranges from 23 to 67 seconds in duration. We obtained 1 to 54 speech samples per participant, 26 on average.

After every recording session, participants are prompted to fill out the DASS-21 [15] questionnaire, consisting of 21 statements, with 7 questions associated with each of the scales of Depression, Anxiety, and Stress. The participants rated each statement from 0 to 3, indicating how much the statement applied to them. The three scales of the DASS-21 provide scores on depression, anxiety, and stress of the individual in the range of 0 to 21; these scores are then multiplied by 2 for consistency with the more detailed DASS-42 scale [15]. Individuals scoring lower than 9, 7, and 15 (out of 42) on the depression, anxiety, and stress scale respectively are considered healthy. 77%, 72%, and 88% of our samples belong to the normal range of depression, anxiety, and stress respectively. We summarized descriptive statistics of DASS-21 scores in our dataset in Table 1.

Table 1. Descriptive statistics the DASS-21 scores in each language.

		Depression	Anxiety	Stress
English	Mean	9.58	8.03	11.22
	Std.	8.29	8.00	8.71
	Min.	0.00	0.00	0.00
	Median	8.00	6.00	10.00
	Max.	36.00	32.00	34.00
Spanish	Mean	5.01	3.98	6.07
	Std.	7.07	6.36	7.74
	Min.	0.00	0.00	0.00
	Median	2.00	2.00	4.00
	Max.	42.00	42.00	42.00
Overall	Mean	5.94	4.79	7.12
	Std.	7.55	6.90	8.20
	Median	4.00	2.00	4.00

3. DISORDER SEVERITY AND PROTOCOL ADHERENCE

3.1. Relationship Among Disorder Severities

We calculated the Spearman Correlation Coefficient among the DASS-21 scores for all three disorders and observed a high positive correlation among them (Table 2). This indicated that our study participants with subthreshold scores in one disorder consistently scored low on all three disorders, while participants with high DASS-21 scores on one disorder were very likely to also score high on the other disorders.

3.2. Relationship Between DASS-21 Score and Adherence to Study Protocol

Different studies revealed the relationship between the severity of psychological disorders and day-to-day routine activities. Wing *et al.* [16] found a moderate positive correlation between depression severity and adherence to treatment components. In our study, we observed that participants with lower DASS-21 scores adhered better to the data collection procedure, designed as a routine activity to be performed twice a week. Figure 1 represents a scatter plot of the DASS-21 score of depression, anxiety, and stress against the number of days since the last DASS-21 upload. Each point in the graph represents a single sample. We can observe that most points are below a score of 12 and less than ten days since the last upload. Additionally, we calculated the Spearman correlation coefficient and found a moderate linear correlation between the participant's adherence to the protocol and the DASS-21 scores (Table 2).

Table 2. Spearman Correlation Coefficients between DASS-21 scores and participants' adherence to the study protocol.

	Depression	Anxiety	Stress	Adherence
Depression	1.00	0.70	0.74	0.39
Anxiety		1.00	0.90	0.44
Stress			1.00	0.48

4. PREDICTING DASS-21 SCORES FROM SPEECH

We trained individual one-dimensional (1D) CNN models on VGG-19 features extracted from spectrograms of the audio samples to predict the severity of each disorder on DASS-21 scale.

4.1. Data Cleaning

We used the Noisereduce [2] algorithm to clean the speech samples. The algorithm computes the spectrogram of a speech signal and estimates a noise threshold for each frequency band of the signal. The threshold is used to compute a mask that filters the noise below the frequency-varying threshold [17].

4.2. Data Partitioning

We divided our data into five non-overlapping folds. In each fold 20% samples were held out for testing. In each training and test partition, we ensured the same ratio of normal and high DASS-21 scores as the original dataset. Within each training fold, we used 20% samples for validation, maintaining the same proportion of normal and high scores as the original dataset. We also ensured that each speech sample appeared in at least one test set, and no speech sample appeared in multiple test sets.

[2] https://github.com/timsainb/noisereduce

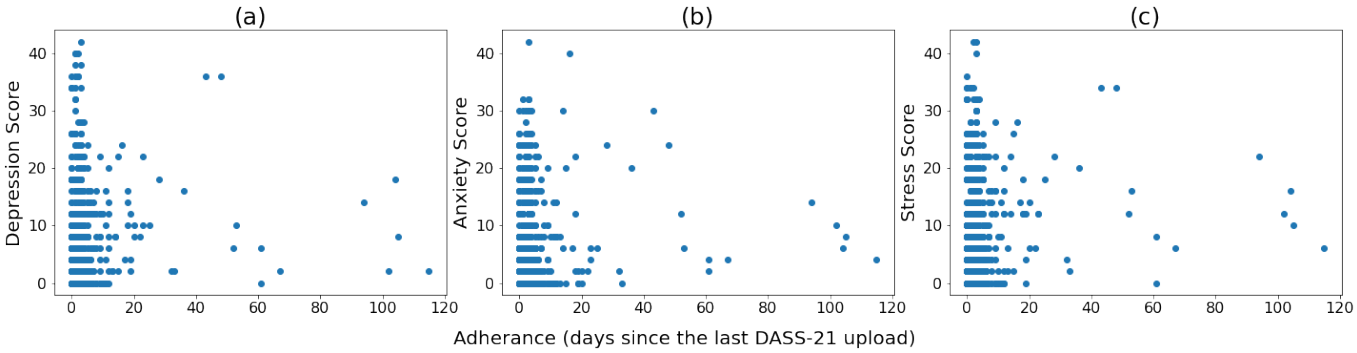

Fig. 1. Scatter plot of the number of days between DASS21 score uploads against DASS-21 scores. Figure (a) represents the depression scores, (b) the anxiety scores and (c) the stress scores.

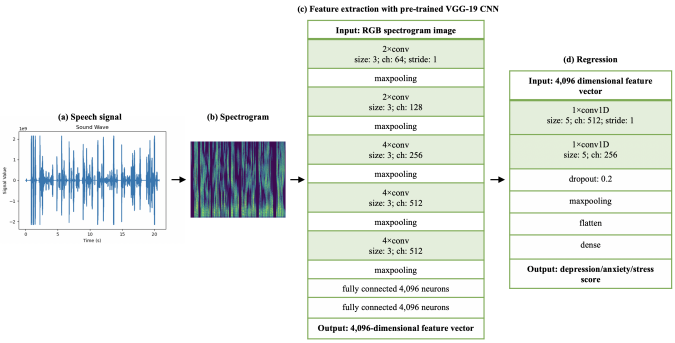

Fig. 2. Our system pipeline. Spectrograms (b) are generated from whole audio files (a) and fed into pre-trained VGG-19 CNN. Activations of the second fully connected layer are extracted as 4,096-dimensional deep spectrum feature vectors (c) used to train a 1-dimensional CNN regressor (d).

4.3. Feature Extraction

We used the DeepSpectrum Python toolkit [3] for feature extraction from the audio samples with pre-trained CNNs. Hamming windows of width 16 ms shifting forward by 8 ms are used to compute the power spectral density on the dB power scale. Matplotlib [4] plots of 387×387 pixels in *viridis* color map are generated, which are then resized to 224×224 pixels to fit the input size of CNN. *Viridis* is a sequential color map varying from blue (lower range) to green to yellow (upper range) (Figure 2(b)).

The spectrograms are then fed into the pre-trained VGG-19 CNN [18]. VGG-19 CNN is a combination of 19 layers including convolutional, maxpooling, and fully connected layers, using rectified linear units (ReLU) as activation functions. To obtain the deep spectrum features, spectrogram plots are forwarded through the pre-trained networks, and the activations from the neurons on the second fully connected layers are extracted as feature vectors. The resulting feature set is a 4,096-dimensional vector, each representing one speech sample. Figure 2(c) illustrates the procedure of extracting VGG-19 features. Figure 2(a) to (d) depicts the complete pipeline of our system.

4.4. Experimental Setting

To predict depression, anxiety, and stress scores we formulated a CNN consisting of two 1D convolutional layers, followed by a dropout layer, a maxpooling layer, and a fully connected layer. ReLU is used as the activation function for the network layers. We used a filter size of 5×1 for the convolutional layers. The dropout rate was 0.2. The stride of the maxpooling was 8. ADAM optimizer was used, setting the learning rate to $10e-5$ with a decay of $10e-7$. We trained the model for 300 epochs in batches of 32 samples. We applied early stopping when the root-mean-square-error (RMSE) loss did not decrease for 20 epochs. The prediction of depression, anxiety, or stress score was obtained from the final fully connected layer. After training on each fold, we obtained the prediction on the test fold. As the test folds collectively contain all the speech samples in our dataset, we report the model performances on the concatenated predictions on the five test folds.

5. RESULTS AND DISCUSSION

We report the performance of our model using the following two metrics:

$$\text{RMSE}(y, \hat{y}) = \sqrt{\frac{\sum_{i=0}^{N-1}(y_i - \hat{y}_i)^2}{N}} \quad (1)$$

Here y and $\hat{y}$ represent the ground truth and predicted scores on ith sample, and N indicates the total number of samples.

$$R^2(y, \hat{y}) = 1 - \frac{(y_i - \hat{y}_i)^2}{(y_i - \overline{y})^2} \quad (2)$$

Where $\overline{y}$ is the mean of the ground truth scores. Table 3 summarized the performance of our regression model in predicting the score of the three disorders.

In our work, we predicted DASS-21 scores for each disorder, ranging from 0 to 42. As the scores in existing literature report their predictions in different scales, we compare

[3] https://github.com/DeepSpectrum/DeepSpectrum
[4] https://matplotlib.org/

Table 3. RMSE and R^2 metrics of the DASS-21 predictions of our proposed CNN model

	RMSE	R^2
Depression	7.09	0.37
Anxiety	7.69	0.47
Stress	8.40	0.36

our performance with the state-of-the-art using the normalized root-mean-square-error (NRMSE) metric calculated as follows:

$$\text{NRMSE} = \frac{RMSE}{y_{max} - y_{min}}\% \tag{3}$$

Here y_{max} and y_{min} are the highest and lowest values on the measurement scale respectively. Table 4 shows the comparison of regression models in existing literature with our proposed model.

Table 4. Comparison of regression models for predicting depression, anxiety, and stress. $(^*)$ indicates multimodal model.

	Citation	**Scale**	**Range**	**NRMSE**
Depression	Kim *et al.*[19]	PHQ-9	27	0.205
	Rodrigues *et al.* [20]	PHQ-9	27	0.211
	He & Cao [21]	BDI-II	63	0.159
	Tasnim & Stroulia [6]	BDI-II	63	0.155
	Tasnim *et al.* [22]	PHQ-8	24	0.221
	Ray *et al.* [23]	PHQ-8	24	0.213
	Our Study	DASS-21	42	0.169
Anxiety	Fatima *et al.** [24]	DASS-21	42	0.089
	Our Study	DASS-21	42	0.183
Stress	Fatima *et al.** [24]	DASS-21	42	0.103
	Our Study	DASS-21	42	0.200

Table 4 shows that our CNN model outperforms most other speech-based models for predicting depression severity and is competitive on the other two disorders, in two languages. There are no other studies predicting anxiety and stress based on acoustic features only. Fatima *et al.* [24] used linguistic features extracted from text data. Our CNN models using acoustic features exclusively predict anxiety and stress with a competitive error ratio. One limitation of our dataset is, that most of our samples fall in the normal range of depression, anxiety, and stress scores respectively, which biases the model's prediction towards subthreshold scores. In our future work, we will consider balancing the dataset by augmenting samples with the higher range of scores. In this work, we considered each sample as an independent instance, as each sample is associated with a DASS-21 score. We plan to explore the possibilities of formulating personalized models by exploiting the longitudinal dataset in our future endeavors.

6. CONCLUSION

In this work, we introduce a new longitudinal and multilingual (English and Spanish) speech corpus for depression, anxiety, and stress. Our dataset captures valuable information on the post-pandemic effect on the mental health of youths. The dataset supports the fact that individuals with lower levels of depression, anxiety, and stress exhibit more conformity with routine activities, demonstrated by a positive correlation between DASS-21 scores and adherence to our data collection protocol. Finally, we propose a CNN model trained on VGG-19 features extracted exclusively from acoustic data recorded in English and Spanish language. In comparison to the state-of-the-art acoustic as well as linguistic models, our proposed model demonstrates competitive performance. The usage of acoustic features exclusively offers two-fold benefits. Firstly, the data does not require going through any transcription, therefore it ensures better privacy of the content of the speech. Secondly, being language-independent in nature, this kind of model extends support to diversified users.

7. REFERENCES

[1] Mitchel Kappen, Marie-Anne Vanderhasselt, and George M Slavich, "Speech as a promising biosignal in precision psychiatry," *Neuroscience & Biobehavioral Reviews*, p. 105121, 2023.

[2] Daniel M Low, Kate H Bentley, and Satrajit S Ghosh, "Automated assessment of psychiatric disorders using speech: A systematic review," *Laryngoscope investigative otolaryngology*, vol. 5, no. 1, pp. 96–116, 2020.

[3] Zehra Shah, Shi-Ang Qi, Fei Wang, Mahtab Farrokh, Mashrura Tasnim, Eleni Stroulia, Russell Greiner, Manos Plitsis, and Athanasios Katsamanis, "Exploring language-agnostic speech representations using domain knowledge for detecting alzheimer's dementia," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–2.

[4] Panagiotis Tzirakis, George Trigeorgis, Mihalis A Nicolaou, Björn W Schuller, and Stefanos Zafeiriou, "End-to-end multimodal emotion recognition using deep neural networks," *IEEE Journal of selected topics in signal processing*, vol. 11, no. 8, pp. 1301–1309, 2017.

[5] Alice Othmani, Daoud Kadoch, Kamil Bentounes, Emna Rejaibi, Romain Alfred, and Abdenour Hadid, "Towards robust deep neural networks for affect and depression recognition from speech," in *Pattern Recognition. ICPR International Workshops and Challenges: Virtual Event, January 10–15, 2021, Proceedings, Part II*. Springer, 2021, pp. 5–19.

[6] Mashrura Tasnim and Eleni Stroulia, "Detecting depression from voice," in *Advances in Artificial Intelligence: 32nd Canadian Conference on Artificial Intelligence, Canadian AI 2019, Kingston, ON, Canada, May 28–31, 2019, Proceedings 32*. Springer, 2019, pp. 472–478.

[7] Brian Diep, Marija Stanojevic, and Jekaterina Novikova, "Multi-modal deep learning system for depression and anxiety detection," in *Empowering Communities: A Participatory Approach to AI for Mental Health*, 2022.

[8] Jonathan Gratch, Ron Artstein, Gale M Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, et al., "The distress analysis interview corpus of human and computer interviews.," in *LREC*. Reykjavik, 2014, pp. 3123–3128.

[9] Michel Valstar, Jonathan Gratch, Björn Schuller, Fabien Ringeval, Denis Lalanne, Mercedes Torres Torres, Stefan Scherer, Giota Stratou, Roddy Cowie, and Maja Pantic, "Avec 2016: Depression, mood, and emotion recognition workshop and challenge," in *Proceedings of the 6th international workshop on audio/visual emotion challenge*, 2016, pp. 3–10.

[10] Fabien Ringeval, Björn Schuller, Michel Valstar, Jonathan Gratch, Roddy Cowie, Stefan Scherer, Sharon Mozgai, Nicholas Cummins, Maximilian Schmitt, and Maja Pantic, "Avec 2017: Real-life depression, and affect recognition workshop and challenge," in *Proceedings of the 7th annual workshop on audio/visual emotion challenge*, 2017, pp. 3–9.

[11] Fabien Ringeval, Björn Schuller, Michel Valstar, Nicholas Cummins, Roddy Cowie, Leili Tavabi, Maximilian Schmitt, Sina Alisamir, Shahin Amiriparian, Eva-Maria Messner, et al., "Avec 2019 workshop and challenge: state-of-mind, detecting depression with ai, and cross-cultural affect recognition," in *Proceedings of the 9th International on Audio/visual Emotion Challenge and Workshop*, 2019, pp. 3–12.

[12] Mashrura Tasnim, Malikeh Ehghaghi, Brian Diep, and Jekaterina Novikova, "Depac: a corpus for depression and anxiety detection from speech," in *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology*, 2022, pp. 1–16.

[13] Damian F Santomauro, Ana M Mantilla Herrera, Jamileh Shadid, Peng Zheng, Charlie Ashbaugh, David M Pigott, Cristiana Abbafati, Christopher Adolph, Joanne O Amlag, Aleksandr Y Aravkin, et al., "Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 due to the covid-19 pandemic," *The Lancet*, vol. 398, no. 10312, pp. 1700–1712, 2021.

[14] Steven Weinberger, "Speech accent archive," *George Mason University*, 2015.

[15] Sydney H Lovibond, "Manual for the depression anxiety stress scales," *Sydney psychology foundation*, 1995.

[16] Rena R Wing, Suzanne Phelan, and Deborah Tate, "The role of adherence in mediating the relationship between depression and health outcomes," *Journal of psychosomatic research*, vol. 53, no. 4, pp. 877–881, 2002.

[17] Tim Sainburg, Marvin Thielk, and Timothy Q Gentner, "Finding, visualizing, and quantifying latent structure across diverse animal vocal repertoires," *PLoS computational biology*, vol. 16, no. 10, pp. e1008228, 2020.

[18] Karen Simonyan and Andrew Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[19] Ah Young Kim, Eun Hye Jang, Seung-Hwan Lee, Kwang-Yeon Choi, Jeon Gue Park, and Hyun-Chool Shin, "Automatic depression detection using smartphone-based text-dependent speech signals: Deep convolutional neural network approach," *Journal of Medical Internet Research*, vol. 25, pp. e34474, 2023.

[20] Mariana Rodrigues Makiuchi, Tifani Warnita, Kuniaki Uto, and Koichi Shinoda, "Multimodal fusion of bert-cnn and gated cnn representations for depression detection," in *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, 2019, pp. 55–63.

[21] Lang He and Cui Cao, "Automated depression analysis using convolutional neural networks from speech," *Journal of biomedical informatics*, vol. 83, pp. 103–111, 2018.

[22] Mashrura Tasnim and Jekaterina Novikova, "Cost-effective models for detecting depression from speech," in *2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 2022, pp. 1687–1694.

[23] Anupama Ray, Siddharth Kumar, Rutvik Reddy, Prerana Mukherjee, and Ritu Garg, "Multi-level attention network using text, audio and video for depression prediction," in *Proceedings of the 9th international on audio/visual emotion challenge and workshop*, 2019, pp. 81–88.

[24] Asra Fatima, Ying Li, Thomas Trenholm Hills, and Massimo Stella, "Dasentimental: Detecting depression, anxiety, and stress in texts via emotional recall, cognitive networks, and machine learning," *Big Data and Cognitive Computing*, vol. 5, no. 4, pp. 77, 2021.