# Multi-modal deep learning system for depression and anxiety detection

**Brian Diep**
Department of Computer Science
University of Toronto
Toronto, ON
bdiep@cs.utoronto.edu [*]

**Marija Stanojevic**
Winterlight Labs
Toronto, ON
marija@winterlightlabs.com

**Jekaterina Novikova**
Winterlight Labs
Toronto, ON
jekaterina@winterlightlabs.com

## Abstract

Traditional screening practices for anxiety and depression pose an impediment to monitoring and treating these conditions effectively. However, recent advances in NLP and speech modelling allow textual, acoustic, and hand-crafted language-based features to jointly form the basis of future mental health screening and condition detection. Speech is a rich and readily available source of insight into an individual's cognitive state and by leveraging different aspects of speech, we can develop new digital biomarkers for depression and anxiety. To this end, we propose a multi-modal system for the screening of depression and anxiety from self-administered speech tasks. The proposed model integrates deep-learned features from audio and text, as well as hand-crafted features that are informed by clinically-validated domain knowledge. We find that augmenting hand-crafted features with deep-learned features improves our overall classification F1 score comparing to a baseline of hand-crafted features alone from 0.58 to 0.63 for depression and from 0.54 to 0.57 for anxiety. The findings of our work suggest that speech-based biomarkers for depression and anxiety hold significant promise in the future of digital health.

## 1 Introduction

Depression and anxiety are two of the most common psychiatric disorders that, depending on their severity, can have a profound impact on an individual's well-being and the quality of life [13, 12, 27, 18, 26]. Thus, it is imperative that treatments for depression and anxiety are prioritized as intervention can greatly improve patient outcomes [6, 25]. Global improvement of anxiety and depression treatment options is estimated to have a direct economic benefit over the period from 2016 to 2030 of $239 billion and $169 billion, respectively [4].

Despite the importance of bettering the treatment pipeline, many barriers remain. One of the primary barriers to effective depression and anxiety treatment is the screening process. Traditional methods for screening have a high burden on clinicians and patients in terms of their ease of administration and scoring, no clear reference standard, and the degree of patient activation and monitoring required [23].

[*] This work was done while at Winterlight Labs

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

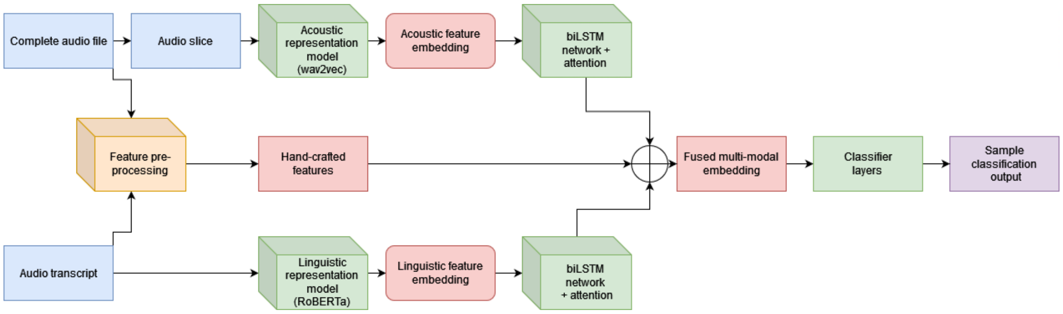
Complete audio file
Audio slice
Acoustic representation model (wav2vec)
Acoustic feature embedding
biLSTM network + attention
Feature pre-processing
Hand-crafted features
Fused multi-modal embedding
Classifier layers
Sample classification output
Audio transcript
Linguistic representation model (RoBERTa)
Linguistic feature embedding
biLSTM network + attention

Figure 1: Classification module architecture diagram

Assessment scales such as the Patient Health Questionnaire (PHQ-8) [20] or Generalized Anxiety Disorder (GAD-7) [29] offer a more quantitative basis for screening.

From another perspective, speech and language are two modalities that form a promising and objective basis for mental health screening. It is well-established that depression and anxiety can alter an individual's general cognition, with specific biases in their attention and memory [5, 22]. These deficits can manifest in altered acoustic and linguistic dimensions of speech. Some of these include altered rate of speech or increased usage of first-person pronouns [24, 16].

With recent advances in natural language processing and computational power, we now have the ability to collect, measure, and analyze speech data on a larger scale. There is also the rise in popularity of digital platforms such as Amazon Mechanical Turk (mTurk) that has eased the burden of data collection from clinically significant populations [8, 30]. All of this has accelerated development of ML models using speech-based biomarkers for depression and anxiety. These include models that classify anxiety and depression as well as those that predict the severity of these diseases [3, 31, 35]. We build upon the existing literature and extend AudiBERT [31] for the classification of depression and anxiety from speech. Our model incorporates more recent sub-module advances in the architecture and experimental settings. Importantly, we also combine both deep-learned and hand-crafted features to best capture the signal of depression and anxiety that is carried through the acoustic and linguistic properties of speech. We demonstrate that our model achieves better performance on the validation dataset.

## 2 Modeling

Depression and anxiety can present themselves through acoustic and linguistic features of speech [24, 16]. Therefore, our architecture (Figure 1) leverages both of these modalities by parallel representation learning from audio and textual data in addition to representation learning from features hand-crafted by domain experts. Our architecture is inspired by AudiBERT [31].

Working with deep-learned representations of speech can allow for our models to capture more abstract signals in speech that can be used for better depression/anxiety detection. In our work, we use pre-trained speech and language representation models which have been shown to be effective and robust for generating representations of acoustics and text [2, 19].

We utilize Wav2Vec 2.0, one of the best acoustic signal representation models, to learn the features from the speech signal. The output of the Wav2Vec 2.0 base-model, pre-trained on 100k hours of the Vox-Populi dataset [34], is forwarded to a two-layer biLSTM [14, 11] and then to a multi-head attention layer with two heads. Vectors $R1$, outputs of multi-head attention representing acoustic signal, are used jointly with linguistic and hand-crafted features for classification.

Transformers-based architectures [7] have significantly improved language representation, and performance on variety of domain-specific tasks including emotion classification [28]. To represent transcripts of human speech, we select the base model of RoBERTa [19], as one of the best performing language models which can be trained with a single GPU. The output of RoBERTa is forwarded to a two-layer biLSTM, whose output is redirected to a multi-head attention layer with two heads. Vectors $R2$ are outputs of multi-head attention representing the linguistic signal.

There is a rich body of work studying the pathology of depression and anxiety that suggests specific changes in the acoustic, the semantic, and lexico-syntactic content of the speech of those who are suffering from these diseases [24, 16]. We use domain-experts hand-crafted features $R3$ as additional signal. List of those features can be found in the Appendix in Table A.1 and A.2.

Vectors $R1$, $R2$, and $R3$ are concatenated together to create a combined representation embedding of the subject's speech. This representation is passed through two feedforward layers followed by a binary cross-entropy loss. The architecture classifies between disease and no disease. We train two different models, one for depression and another for anxiety task.

# 3 Experimental setup

We train and evaluate our models using 5-fold cross-validation with the folds constructed such that there is no overlap between the subjects in the training and test fold. We report the mean of precision, recall and F1-score for each model over the 5 folds. Results are achieved using AdamW optimization with learning rate $lr = 3e - 5$. We use binary cross-entropy with logits loss from the PyTorch library. The model is trained on T4 Tensor Core GPU with 16 GB RAM.

We use Wav2Vec 2.0 and RoBERTa implementations from the *HuggingFace* library. Due to memory and architecture constraints with inputting large audio files into Wav2Vec2, we also split the audio samples into consecutive 10 second intervals. The audio was sampled at a rate of 16 000 Hz. Then Wav2Vec2 feature extractor is used to create the input. RoBERTa's input is a speech transcript generated from the audio via ASR. The text is further transformed by the RoBERTa tokenizer and padded to length of 512. Note, we also add several tokens to the tokenizer corresponding to a set of unfilled and filled pauses in the speech. Pre-trained model weights are not frozen and are fine-tuned for 10 epochs with a batch size of 4 due to GPU memory constraints.

As a baseline, we train a feedforward network using hand-crafted features provided by domain-experts only. The network consists of five linear layers followed by Leaky ReLU activation function [21]. Every layer is twice smaller than the previous one and we use a dropout of 0.2 throughout the network. Network is trained using AdamW optimization with learning rate $lr = 3e - 4$, batch size of 8, and binary cross-entropy with logits loss. We use the same 5-fold cross-validation process as for the proposed model and we report the mean of precision, recall and F1-score.

## 3.1 Dataset

The dataset used to train and test the model comes from an extended version of the DEPAC corpus [30]. The DEPAC corpus contains crowd-sourced (mTurk) audio samples from 3543 unique individuals performing a range of self-administered speech tasks. For the purposes of this analysis, we subset the data to only include speech from the tasks that contain elements of narrative speech. In total, the dataset contains 4209 unique audio samples and corresponding audio transcripts from the below-mentioned speech tasks.

**Journaling** and **prompted narrative tasks:** the participant is asked to describe an experience or event based on a given prompt. For journaling task, they are asked about their day whereas in prompted narrative, they are also asked about hobbies or travel experiences depending on the specific prompt. These narrative speech tasks can contain signals relevant for depression or anxiety prediction [32].

**Semantic fluency task:** the participant is prompted to describe within one minute positive experiences that will occur in the future. Similar verbal fluency tasks have been shown to correlate with issues with executive function associated with depression [9].

The dataset contains the self-rated PHQ-8 and GAD-7 scores for each individual. GAD-7 is rated on a scale of 0-21 and PHQ-8 on a scale of 0-24. Following AudiBERT, literature [20, 29], and consultations with experts, we adopt binary classification tasks. We convert these scores into a "soft" binary diagnosis label using a score of 10 as a cutoff on both scales. Approximately 25.3% of subjects had a PHQ-8 score above 9, and 12.8% had a GAD-7 score above 9 (diagnosis).

For each complete audio sample and transcript, we extract the hand-crafted features whose list is given in the Appendix A.1 and A.2.

Table 1: Anxiety and depression classification results. Bold indicates highest F1 score per disease.

| | Anxiety | | | | | | Depression | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Hand-crafted features only | | | Deep-learned + hand-crafted features | | | Hand-crafted features only | | | Deep-learned + hand-crafted features | | |
| | Precision | Recall | F1 | Precision | Recall | F1 | Precision | Recall | F1 | Precision | Recall | F1 |
| No diagnosis ( score<10) | 0.81 | 0.65 | 0.72 | 0.76 | 0.72 | **0.73** | 0.73 | 0.78 | 0.75 | 0.77 | 0.83 | **0.80** |
| Diagnosis (score≥ 10) | 0.28 | 0.41 | 0.33 | 0.37 | 0.42 | **0.40** | 0.31 | 0.42 | 0.35 | 0.48 | 0.39 | **0.43** |
| Overall | | | 0.54 | | | **0.57** | | | 0.58 | | | **0.63** |

## 4 Results and Discussion

The results of our experiments are displayed in Table 1. Examining them in aggregate reveals that our models perform better in predicting no diagnosis, and they are struggling to predict diagnosis. We hypothesize that this is partially a function of the data imbalance that exists within our dataset, as most collected depression and anxiety data comes from individuals with lower scores.

The results show that the inclusion of deep-learned features enriches the representation by adding properties that are not fully captured the hand-crafted features, improving the detection of depression and anxiety. This reflects previous results [31], where the addition of deep-learned features, especially text representation models, improved classification performance for depression.

One of the challenges with developing models for classification of depression and anxiety comes from the distribution of data. In our data and much corpora, a majority of the subjects ware classified with having PHQ-8/GAD-7 scores under 10 leading to class imbalance [33, 10]. Imbalance in classes in training data poses a hurdle in development of robust models [17]. Furthermore, within the classes, the distribution of scores is still uneven. A distribution of PHQ-8/GAD-7 scores is long-tailed and skewed towards lower severity cases. This can lead to issues of within-class imbalance that are difficult to resolve [15].

Interestingly, we also find that depression classification results in higher overall F1-score than anxiety classification. One reason for this was likely due to the data imbalance issue in anxiety samples, which was particularly pronounced as compared to depression (12.8% vs. 25.3% with scores above 9). Another potential reason for this worse performance is that acoustic features in anxiety have been shown to not vary as much with severity as compared with depression [1]. This suggests that anxiety prediction through speech-assessment is a harder task than its corollary in depression.

These findings add to the existing body of work that speech is an appropriate modality for depression and anxiety biomarker development. In particular, using both hand-crafted and deep-learned features maximizes the signal that can be extracted from the speech stream. It also shows how prediction performance for these models is often variable with respect to anxiety/depression severity.

## 5 Conclusion

In this work, we present a model for the prediction of anxiety and depression from self-administered speech tasks. Our models extend upon previous work that focuses on classification of depression and anxiety and combines it with a set of hand-crafted features that is able to capture many of the nuanced changes in acoustic and linguistic content of depressed and anxious speech. We find that the proposed model, that combines hand-crafted features with deep-learning speech and language representation, improves classification F1-score of both classes compared to the baseline. The results presented in this paper form a promising basis towards the development of better screening tools for anxiety and depression via speech data.

## References

[1] Luciana Albuquerque, Ana Rita S Valente, António Teixeira, Daniela Figueiredo, Pedro Sa-Couto, and Catarina Oliveira. Association between acoustic speech features and non-severe levels of anxiety and depression symptoms across lifespan. *PloS one*, 16(4):e0248842, 2021.

[2] Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations, 2020. URL `https://arxiv.org/abs/2006.11477`.

[3] Tathagata Banerjee, Matthew Kollada, Pablo Gersberg, Oscar Rodriguez, Jane Tiller, Andrew E Jaffe, and John Reynders. Predicting mood disorder symptoms with remotely collected videos using an interpretable multimodal dynamic attention fusion network. *arXiv preprint arXiv:2109.03029*, 2021.

[4] Dan Chisholm, Kim Sweeny, Peter Sheehan, Bruce Rasmussen, Filip Smit, Pim Cuijpers, and Shekhar Saxena. Scaling-up treatment of depression and anxiety: a global return on investment analysis. *The Lancet Psychiatry*, 3(5):415–424, 2016. ISSN 2215-0366. doi: https://doi.org/10.1016/S2215-0366(16)30024-4. URL `https://www.sciencedirect.com/science/article/pii/S2215036616300244`.

[5] Robert M Cohen, Herbert Weingartner, Sheila A Smallberg, David Pickar, and Dennis L Murphy. Effort and cognition in depression. *Archives of general psychiatry*, 39(5):593–597, 1982.

[6] Mark R Dadds, Susan H Spence, Denise E Holland, Paula M Barrett, and Kristin R Laurens. Prevention and early intervention for anxiety disorders: a controlled trial. *Journal of Consulting and Clinical Psychology*, 65(4):627, 1997.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2018. URL `https://arxiv.org/abs/1810.04805`.

[8] Krista Engle, Margaret Talbot, and Kristin W Samuelson. Is amazon's mechanical turk (mturk) a comparable recruitment source for trauma studies? *Psychological Trauma: Theory, Research, Practice, and Policy*, 12(4):381, 2020.

[9] Philippe Fossati, Anne-Marie Ergis, Jean-François Allilaire, et al. Qualitative analysis of verbal fluency in depression. *Psychiatry research*, 117(1):17–24, 2003.

[10] Jonathan Gratch, Ron Artstein, Gale 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. Technical report, UNIVERSITY OF SOUTHERN CALIFORNIA LOS ANGELES, 2014.

[11] Alex Graves and Jürgen Schmidhuber. Framewise phoneme classification with bidirectional lstm and other neural network architectures. *Neural networks*, 18(5-6):602–610, 2005.

[12] Barry Gurland. The impact of depression on quality of life of the elderly. *Clinics in geriatric medicine*, 8(2):377–386, 1992.

[13] Eric R Henning, Cynthia L Turk, Douglas S Mennin, David M Fresco, and Richard G Heimberg. Impairment and quality of life in individuals with generalized anxiety disorder. *Depression and anxiety*, 24(5):342–349, 2007.

[14] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997.

[15] Nathalie Japkowicz. Concept-learning in the presence of between-class and within-class imbalances. In *Conference of the Canadian society for computational studies of intelligence*, pages 67–77. Springer, 2001.

[16] Doerte U Junghaenel, Joshua M Smyth, and Laura Santner. Linguistic dimensions of psychopathology: A quantitative analysis. *Journal of Social and Clinical Psychology*, 27(1):36, 2008.

[17] Bartosz Krawczyk. Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, 5(4):221–232, 2016.

[18] Jean-Pierre Lépine. The epidemiology of anxiety disorders: prevalence and societal costs. *Journal of Clinical Psychiatry*, 63:4–8, 2002.

[19] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[20] Bernd Löwe, Jürgen Unützer, Christopher M Callahan, Anthony J Perkins, and Kurt Kroenke. Monitoring depression treatment outcomes with the patient health questionnaire-9. *Medical care*, pages 1194–1201, 2004.

[21] Andrew L Maas, Awni Y Hannun, Andrew Y Ng, et al. Rectifier nonlinearities improve neural network acoustic models. In *Proc. icml*, volume 30, page 3. Citeseer, 2013.

[22] Andrew Mathews and Colin MacLeod. Cognitive vulnerability to emotional disorders. *Annu. Rev. Clin. Psychol.*, 1:167–195, 2005.

[23] Donald E Nease and Jean M Malouin. Depression screening: a practical strategy. *Journal of Family Practice*, 52(2):118–126, 2003.

[24] Benjamin Pope, Thomas Blass, Aron W Siegman, and Jack Raher. Anxiety and depression in speech. *Journal of Consulting and Clinical Psychology*, 35(1p1):128, 1970.

[25] Charles F Reynolds III, Pim Cuijpers, Vikram Patel, Alex Cohen, Amit Dias, Neerja Chowdhary, Olivia I Okereke, Mary Amanda Dew, Stewart J Anderson, Sati Mazumdar, et al. Early intervention to reduce the global health and economic burden of major depression in older adults. *Annual review of public health*, 33:123–135, 2012.

[26] Derek Richards. Prevalence and clinical course of depression: a review. *Clinical psychology review*, 31(7):1117–1125, 2011.

[27] Babak Roshanaei-Moghaddam, Wayne J Katon, and Joan Russo. The longitudinal effects of depression on physical activity. *General hospital psychiatry*, 31(4):306–315, 2009.

[28] Shamane Siriwardhana, Andrew Reis, Rivindu Weerasekera, and Suranga Nanayakkara. Jointly fine-tuning" bert-like" self supervised models to improve multimodal speech emotion recognition. *arXiv preprint arXiv:2008.06682*, 2020.

[29] Robert L Spitzer, Kurt Kroenke, Janet BW Williams, and Bernd Löwe. A brief measure for assessing generalized anxiety disorder: the gad-7. *Archives of internal medicine*, 166(10): 1092–1097, 2006.

[30] 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*, pages 1–16, 2022.

[31] Ermal Toto, ML Tlachac, and Elke A Rundensteiner. Audibert: A deep transfer learning multimodal classification framework for depression screening. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 4145–4154, 2021.

[32] Raluca Nicoleta Trifu, Bogdan NEMEŞ, Carolina Bodea-Haţegan, and Doina Cozman. Linguistic indicators of language in major depressive disorder (mdd). an evidence based research. *Journal of Evidence-Based Psychotherapies*, 17(1), 2017.

[33] Michel Valstar, Björn Schuller, Kirsty Smith, Timur Almaev, Florian Eyben, Jarek Krajewski, Roddy Cowie, and Maja Pantic. Avec 2014: 3d dimensional affect and depression recognition challenge. In *Proceedings of the 4th international workshop on audio/visual emotion challenge*, pages 3–10, 2014.

[34] Changhan Wang, Morgane Rivière, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation, 2021. URL `https://arxiv.org/abs/2101.00390`.

[35] Le Yang, Dongmei Jiang, Xiaohan Xia, Ercheng Pei, Meshia Cédric Oveneke, and Hichem Sahli. Multimodal measurement of depression using deep learning models. In *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, pages 53–59, 2017.

# A Appendix

Table A.1: Summary of the linguistic features.

| **Feature Group** | **Motivations** |
|---|---|
| Discourse mapping | Techniques to formally quantify utterance similarity and disordered speech via distance metrics or graph-based representations. |
| Local coherence | Coherence and cohesion in speech is associated with the ability to sustain attention and executive functions. |
| Lexical complexity and richness | Language pattern changes in particular related to the irregular usage patterns of words of certain grammatical categories. |
| Syntactic complexity | Measures of syntactic complexity of utterances. |
| Utterance cohesion | Measures of tense and concordance within utterances. |
| Sentiment | Features such as valence, arousal, and dominance. |
| Word finding difficulty | Metrics related to disfluency and filled pauses in speech. |

Table A.2: Summary of the acoustic features.

| **Feature Group** | **Motivations** |
|---|---|
| Intensity (auditory model based) | Perceived loudness in $dB$ relative to normative human auditory threshold. |
| MFCC 0-12 | MFCC 0-12 and energy, their first and second order derivatives are calculated on every 16 ms window and step size of 8 ms, and then, averaged over the entire sample. |
| Zero-crossing rate (ZCR) | Zero crossing rate across all the voiced frames showing how intensely the voice was uttered. |
| $F_0$ | Fundamental frequency in Hz. |
| Harmonics-to-noise-ratio (HNR) | Degree of acoustic periodicity. |
| Jitter and shimmer | Jitter is the period perturbation quotient and shimmer is the amplitude perturbation quotient representing the variations in the fundamental frequency. |
| Durational features | Total audio and speech duration in the sample. |
| Pauses and fillers | Number and duration of short ($< 1s$), medium ($1-2s$) and long ($> 2s$) pauses, mean pause duration, and pause-to-speech ratio. |
| Phonation rate | Number of voiced time windows over the total number of time windows in a sample. |