INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

# Bayesian Networks for the robust and unbiased prediction of depression and its symptoms utilizing speech and multimodal data

*Salvatore Fara*[1], *Orlaith Hickey*[1], *Alexandra Georgescu*[1,2], *Stefano Goria* [1], *Emilia Molimpakis*[1], *Nicholas Cummins*[1,2]

[1]Thymia, London, UK
[2]Institute of Psychiatry, Psychology & Neuroscience (IoPPN), King's College London, London, UK
{salvatore,orlaith,alexandra,stefano,emilia,nick}@thymia.ai, nick.cummins@kcl.ac.uk

## Abstract

Predicting the presence of major depressive disorder (MDD) using speech is highly non-trivial. The heterogeneous clinical profile of MDD means that any given speech pattern may be associated with a unique combination of depressive symptoms. Conventional discriminative machine learning models may lack the complexity to robustly model this heterogeneity. Bayesian networks, however, are well-suited to such a scenario. They provide further advantages over standard discriminative modeling by offering the possibility to (i) fuse with other data streams; (ii) incorporate expert opinion into the models; (iii) generate explainable model predictions, inform about the uncertainty of predictions, and (iv) handle missing data. In this study, we apply a Bayesian framework to capture the relationships between depression, depression symptoms, and features derived from speech, facial expression and cognitive game data. Presented results also highlight our model is not subject to demographic biases.

**Index Terms**: Depression, Bayesian Networks, Fusion, Knowledge Integration, Missing Data, Fairness

## 1. Introduction

The healthcare sector is in urgent need of better tools to tackle the challenges of major depressive disorder (MDD) efficiently and effectively. Depression assessments are still based on self-report questionnaires which are prone to bias [1] and where the variability between individuals' interpretation of questionnaire items is high [2]. Furthermore, clinical interviews and observation are naturally influenced by the clinician's experience and acumen [3]. Collectively, this means identifying the correct diagnosis and treatment can take many years, with some studies finding untreated depression rates as high as 77% [4]. There is an immediate need for a clinical decision support tool offering objective depression metrics, as easily accessible and reliably trackable as physical health ones (e.g. blood test markers). Advances in digital health and phenotyping technologies are therefore being considered integral to improving MDD-associated clinical pathways [4].

In recent years, there has been an acceleration in the number of papers centred around the application of *machine learning* in the domain of digital health. These works include analyses of speech, facial expressions and cognitive assessments to provide objective measurement criteria to aid in MDD diagnosis [5, 6]. A potential shortcoming of such works, however, is that they have almost exclusively been focused on supervised modelling paradigms learning how to partition data based on subjective depression scales, such as the 8-item Patient Health Questionnaire (PHQ-8; [7]), thereby also becoming subject to the same concerns around self-report subjectivity. Moreover, they mainly utilise large multivariate feature spaces and deep learning models which lack transparency regarding how their decisions are being made [8]. Alongside this lack of explainability, such approaches also lack the ability to incorporate expert opinion into the model and are unable to handle missing data robustly.

Bayesian Networks (BN) offer a natural framework to satisfy all the above requirements, which are common in healthcare modelling. Indeed, a few recent works have successfully adopted BNs to tackle mental health modelling problems; for example, [9, 10]. However, the predictors in these approaches have been simple demographics, biological or environmental factors, as opposed to rich multimodal datasets that can include audio and video data. Only a very small number of works have explored the use of BNs for detecting depression from speech e. g. [11, 12]. These works, however, are focused on the classification of depression severity; they do not consider joint classification with symptoms or the inclusion and effects of confounding factors.

In this study we propose a novel BN model for joint MDD and depression symptoms classification given a multimodal feature set containing speech, facial expression and cognitive game data gathered at thymia [13]. We then present a range of experiments demonstrating the model's performance under different realistic use case scenarios, including varying degrees of missing data and integration of expert knowledge.

The main novel contribution of this work is a methodology for incorporating speech and video data in a BN model that achieves strong performance in MDD classification. We also highlight its potential as a clinical decision-support tool by providing results for individual core MDD symptoms and give a detailed breakdown of performance according to key sociodemographic factors.

## 2. Experimental Corpora

This section describes the collection and preprocessing of the data used in our expirements.

### 2.1. Dataset

We trained and validated our models on our data collected in-house. This data collection received ethical approval from the Association of Research Managers and Administrators. Publicly available speech-depression datasets such as the Audio-Visual Depressive (AViD) corpus [14] and Distress Analysis Interview Corpus (DAIC) [15] do not have the required meta-data to support our stated analytical aims. Further, these data have been in the public domain for 10 years so subject to concerns relating to overfitting and multiple hypothesis testing.

The experimental data used in this study consist of 1,336 English-speaking participants who performed a series of short


10.21437/Interspeech.2023-1709

Table 1: *Sociodemographic, depression and activity distributions in the experimental data.*

| Group | Gender | | Age | | Country | | Device | | Cumulative Activity Time | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | **Male** | **Female** | **$< 36$** | **$\geq 36$** | **UK** | **US** | **Phone** | **PC** | **Paragraph** | **Image** | **n-Back** |
| PHQ-8 $\geq 10$ | 135 | 220 | 215 | 140 | 202 | 153 | 19 | 336 | 4:09:55 | 5:23:22 | "1 day, 19:15:27" |
| PHQ-8 $< 10$ | 496 | 485 | 574 | 407 | 424 | 557 | 42 | 938 | 14:15:52 | 19:10:56 | "5 days, 21:20:47" |

online activities within a single session on our Research Platform [13] [1] using their own personal devices (Table 1). We previously presented a portion of this dataset with a smaller number of participants and a focus on audio (acoustic, prosodic and linguistic) and cognitive data gathered from two thymia activities, i.e. an Image Description Task and the n-Back Task [16], as well as individual PHQ-8 items; see [13]. In the present study, we expand the number of data modalities to include video data recorded during the Image Description Task, as well as additional audio data gathered during a Paragraph Reading Task. Additionally, we include information on the type of personal device that was used to perform the activities.

### 2.2. Data Availability

Due to licensing and IP considerations, we are not at this moment making our dataset generally publicly available. However, we are open to partnering with research institutes and individual academics including data sharing upon request.

### 2.3. Data Collection Activities

We focus on data gathered through three data collection activities: the n-Back Task, the Image Description Task and the Paragraph Reading Task. The first two activities have been previously described in detail in [13]. The Paragraph Reading Task required participants to read aloud a short story (Aesop fable "The North Wind and the Sun" widely used within phonetics [17]) while their voice was being recorded via their device's microphone. Herein we abbreviate the activity names to "*n-Back*", "*Image*", and "*Paragraph*".

### 2.4. Data Selection

A total of 1,898 participants enrolled to the study. Participants were excluded from the dataset if either of these applied: (1) they did not complete all three data collection activities; (2) their recordings were corrupted by technical problems (camera/mic malfunctioning); (3) they did not comply to the tasks (did not speak in the speech tasks). On the basis of these criteria, 1,336 participants were selected.

### 2.5. Data Preprocessing

Audio recordings from the speech eliciting activities were converted to single-channel wave files at 16kHz sampling rate using FFmpeg software. Speech tokens were then extracted from the audio files using Amazon Web Services (AWS) Speech-To-Text service Amazon Transcribe.

[1] The thymia Research Platform allows the hosting of complex, remote, multimodal studies on a smart device. During various activities (e.g. questionnaires, cognitive games etc.), data from the device's camera, keyboard, mouse/trackpad and/or touch screen can be streamed to a secure backend. The platform is fully HIPAA-compliant, 2018 EU GDPR-compliant, is ISO27001-certified and NHS Toolkit-compliant.

### 2.6. Features

Data from the three activities was processed to extract a total of 322 features which included: 8 cognitive features (n-Back), 97 video features (Image), 88 extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) acoustic and prosodic features [18] (from both Image and Paragraph), 24 linguistic features (Image), as well as an additional curated set of 17 fine-grained acoustic features (Paragraph). Details on the cognitive, eGeMAPS and speech features were previously provided and can be found in [13].

The video features were extracted using Visage Technologies Software. The software extracted features related to facial translation, rotation and gaze in the 3D space, action units and emotions. An estimated face scale was also provided to normalise values, allowing for changes in a participant's distance from the screen and camera.

The fine-grained acoustic features consist of summary statistics of the formant trajectories extracted from specific voiced audio segments of the Paragraph audio recordings. We used audio segments corresponding to three sets of words chosen to isolate the following vowel sounds [19]: /i/ ('wind', 'which', 'he', 'his', 'him'), /u/ ('should', 'could', 'took', 'two'), /a/ ('hard', 'last', 'and', 'at').

## 3. Bayesian Network Model

### 3.1. Model Definition

Bayesian Networks (BNs) are probabilistic graphical models that specify the joint distribution by defining a set of conditional independence rules that can be easily mapped to a Directed Acyclical Graph (DAG) [20]. Our BN model is composed of four groups of variables: CONFOUNDS, CONDITION, SYMPTOMS and ACTIVITY measures (Figure 1).

The CONFOUNDS group includes age and gender as demographic variables, and a third variable indicating the type of personal device used by the participant to perform the session on our Research Platform. The age variable is modelled as a categorical distribution with four categories representing four age groups (i.e. 18-25, 26-35, 36-45, 46-100), while both gender and device are modelled as Bernoulli distributions with categories 'male'/'female' and 'smartphone'/'PC' respectively. In the following we use $A \in \{0, 1, 2, 3\}$, $G \in \{0, 1\}$ and $D \in \{0, 1\}$ to indicate the age group, gender and device respectively.

The CONDITION variable indicates the presence (PHQ-8 $\geq$ 10) or absence (PHQ-8 $< 10$) of depression. To capture the variation of depression incidence across age groups and genders, we model the condition $C$ as a Bernoulli distribution:

$$p(C|A, G) = Ber(C|logistic(f_c(A, G))) \quad (1)$$

where

$$f_c(A, G) = \omega_{c,0} + \omega_{c,a}A + \omega_{c,g}G. \quad (2)$$

The SYMPTOM variables represent the individual PHQ-8 items. In order to simplify the model, each symptom is converted from its original 4-point scale to a binary scale indicating

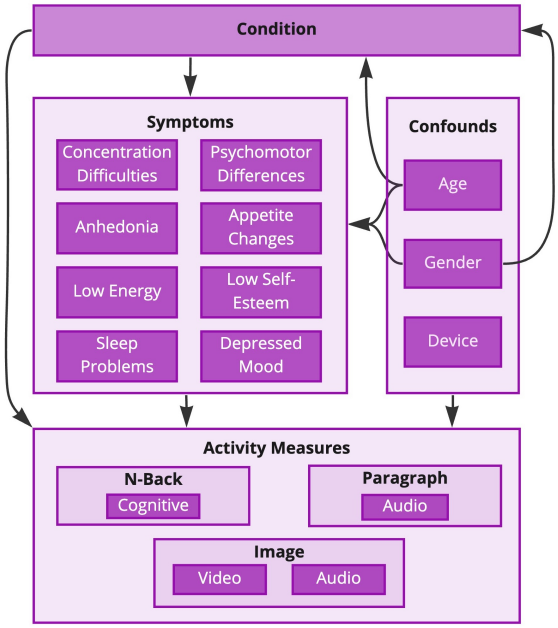


Figure 1: *Overview of the proposed multimodal Bayesian network model.*

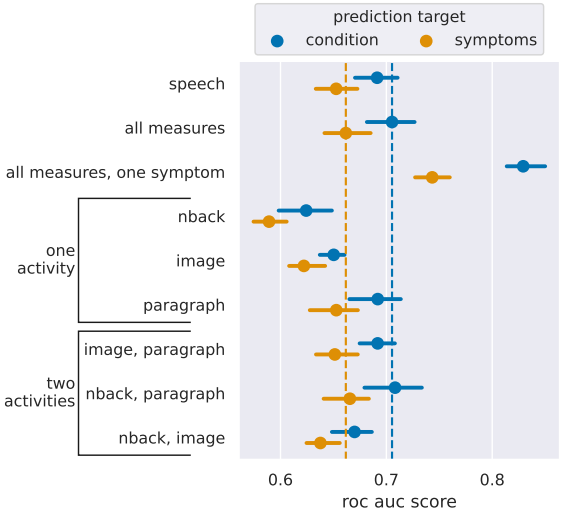


Figure 2: *BN model performance in the joint prediction of condition and symptoms given the observation of other sets of variables. Averages (with 95% CI) across 5 cross-validation test folds are shown. Dashed lines highlight model performances when all measures are observed.*

'low' and 'high' symptom levels. The symptom-specific binarisation thresholds are calculated from a logistic regression of each individual symptom on the condition variable. In order to capture inter-symptom conditional dependencies, the symptom variables are embedded in an inter-symptom DAG estimated using the DirectLiNGAM graph discovery algorithm [21]. Each binary symptom variable $S_s$, with $s \in \{0, 1, ..., 7\}$, is modelled as a Bernoulli distribution:

$$p(S_s|A, G, C, \boldsymbol{P_s}) = Ber(S_s|logistic(f_s(A, G, C, \boldsymbol{P_s}))) \tag{3}$$

where variables in bold are vectors, $\boldsymbol{P_s}$ is a column vector of $k_s$ parent symptoms of $S_s$ as specified by the inter-symptom DAG and

$$f_s(A, G, C, \boldsymbol{P_s}) = \omega_{s,0} + \omega_{s,a}A + \omega_{s,g}G + \omega_{s,c}C + \boldsymbol{\omega_{s,p}}\boldsymbol{P_s} \tag{4}$$

with $\boldsymbol{\omega_{s,p}} \in \mathbb{R}^{k_s}$. In the following we use $\boldsymbol{S}$ to indicate the column vector of all binary symptoms.

The ACTIVITY measures are derived from the feature sets described in the previous section by applying two processing steps. First, standard rescaling is applied to all features individually. Second, supervised PCA [22] is applied to each feature set independently using the condition variable as target. The first two principal components of each feature set are then selected, yielding a total of 16 activity measures (2 from N-Back, 10 from Image, 4 from Paragraph). Each activity measure variable $M_m$, with $m \in \{0, 1, ..., 15\}$, is modelled as a Gaussian distribution

$$p(M_m|A, G, D, C, \boldsymbol{S}) = \mathcal{N}(M_m|f_m(A, G, D, C, \boldsymbol{S}), \sigma_m^2) \tag{5}$$

where

$$f_m(A, G, D, C, \boldsymbol{S}) = \omega_{m,0} + \omega_{m,a}A + \omega_{m,g}G + \omega_{m,d}D + \omega_{m,c}C + \boldsymbol{\omega_{m,s}}\boldsymbol{S} \tag{6}$$

with $\boldsymbol{\omega_{m,s}} \in \mathbb{R}^8$. In the following we use $\boldsymbol{M}$ to indicate the column vector of all activity measures.

The full BN model describing the joint probability distribution over all the variables described above is then given by

$$\begin{aligned} p(\boldsymbol{M}, \boldsymbol{S}, C, A, G, D) = & \prod_{m=0}^{15} p(M_m|A, G, D, C, \boldsymbol{S}) \times \\ & \prod_{s=0}^{7} p(S_s|A, G, C, \boldsymbol{P_s}) \times \\ & p(C|A, G) \times \\ & p(A) \times p(G) \times p(D). \end{aligned} \tag{7}$$

### 3.2. Model Implementation and Training

We implemented the BN model using the probabilistic programming library NumPyro (version 0.11.0) [23] and Python 3.9.15. Model training was performed via the Markov Chain Monte Carlo (MCMC) inference of model parameters using the No-U-Turn sampler (NUTS) algorithm in NumPyro, with 4 Markov chains and 1000 samples per chain. We used the following prior distributions for the model parameters: $Dirichlet(K = 4, \alpha = 1)$ for the group frequencies of the age variable; $Beta(\alpha = 1, \beta = 1)$ for the Bernoulli probabilities of the gender and device variables; $\mathcal{N}(\mu = 0, \sigma = 1)$ for all $\omega$ parameters in Equations (2), (4) and (6); $LogNormal(\mu = 0, \sigma = 1)$ for the $\sigma_m$ parameters in Equation (5).

### 3.3. Model Evaluation

We performed a stratified 5-fold cross-validation to evaluate model performance. The same proportion of gender, age groups and PHQ-8 distribution was kept across the training and test sets for each fold. We use the area under the receiver operating characteristic curve (ROC-AUC) as our evaluation metric.

Table 2: *Mean and SD AUC of our BN and RF models. For BN we add results when slicing the data for gender, age and country.*

| Model | Population | Target | | | |
|---|---|---|---|---|---|
| | | Condition | | Symptoms | |
| | | Mean | SD | Mean | SD |
| BN | Overall | 0.705 | 0.029 | 0.662 | 0.026 |
| | Female | 0.700 | 0.055 | 0.668 | 0.039 |
| | Male | 0.705 | 0.042 | 0.648 | 0.027 |
| | UK | 0.660 | 0.075 | 0.613 | 0.050 |
| | US | 0.746 | 0.019 | 0.707 | 0.027 |
| | Age $< 36$ | 0.693 | 0.036 | 0.644 | 0.034 |
| | Age $\geq 36$ | 0.726 | 0.034 | 0.687 | 0.017 |
| RF | Overall | 0.714 | 0.027 | 0.665 | 0.028 |

### 3.4. Benchmark model

In order to benchmark the BN model, we additionally trained a multi-output Random Forest classifier (RF) which received the same 16 input activity measures used by the BN model. The model had 500 trees with maximum depth of 5 and was implemented using the scikit-learn package (version 1.0) [24]. The maximum depth was set via grid search over [1, 5, 10, 15, 20] using a 5-fold cross-validation.

## 4. Results and Discussions

Given the generative nature of a BN model, any of its variables or groups of variables can be chosen as targets in a prediction task. Of particular interest is the task of predicting CONDITION and SYMPTOMS given the observation of other variables in the model. We performed a set of experiments to evaluate the model performance in this joint prediction task under several realistic scenarios (Figure 2). Overall, the experiments showed an increase in predictive performance with the amount of observed variables in the model, and a generally higher performance for CONDITION than SYMPTOMS. When all ACTIVITY measures are available, the average ROC-AUC is above 0.7 for CONDITION and 0.66 for SYMPTOMS.

Additionally, we evaluated the performance when only subsets of ACTIVITY measures are available as input to the model (Figure 2). These experiments correspond to common real-life scenarios in which a participant does not perform the full set of activities or opts out of recording. This set of experiments revealed an increase in model performance as more activities are observed, with paragraph measures having the strongest positive impact.

Finally, we performed a set of experiments where all ACTIVITY measures plus one SYMPTOM are observed (Figure 2). This simulates the scenario in which reliable information about the presence or absence of a symptom is available to the clinician using the model. In this scenario, we observed that the predictive performance further improves for both CONDITION and the other unknown SYMPTOMS.

To assess the robustness of the model for potential demographics biases, we also performed a segmentation of the performance metrics across gender, age and country (Table 2). The small differences across the demographics splits we considered suggest that our results are not biased.

When benchmarking our BN model against a RF classifier, we can see that both have similar performances (Table 2). This could be viewed as a limitation of our model; however, it is worth considering that we needed only a single BN to perform multiple predictions, whereas separate RF models need to be trained for each inference task. Furthermore, although the performance of our BN model is below that of other approaches in the literature, we believe that our results are a true reflection of what could be expected from a dataset of this size once biases have been minimised [25].

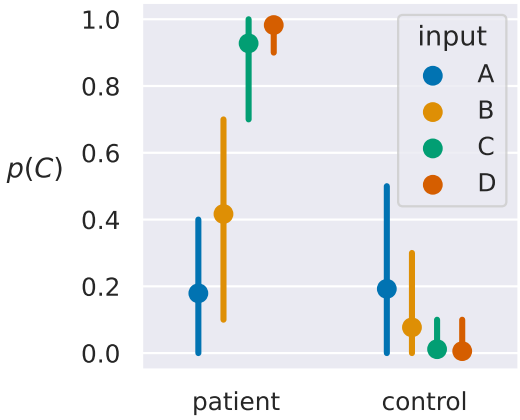


Figure 3: *Raw model predictions of condition probability in two sample participants, one patient and one control, for four sets of inputs: A = confounds, B = confounds + n-back, C = confounds + n-back + paragraph, D = confounds + n-back + paragraph + sleep symptom. Error bars denote 95% credible intervals.*

Finally, the main purpose of our BN model is to serve as a support tool for clinical decision-making. The model allows for (i) the integration of multimodal information alongside speech; and (ii) the clinician using it to provide their expert knowledge. Given its ability to generate predictions despite missing information, this model naturally lends itself to be used as part of an iterative screening process. For example, the clinician may decide to administer only a subset of activities to a patient, then consult the model predictions and decide whether other information may be needed to support a diagnosis, subsequently administering additional activities, asking the patient about their sleep patterns or investigating other symptoms (Figure 3).

## 5. Conclusions

This work represents a proof-of-concept for validating a BN model, demonstrating its performance as a robust speech-based MDD prediction tool. We also highlight the potential of this model under different real-world operating conditions and assess it for potential biases in core sociodemographic factors. A limitation of the current model is the reduced set of confounding variables. Future research will explore a larger set of confounds, such as life events that could affect mood (e.g. loss or change of jobs) or health problems that could affect voice (e.g. having a cold). An additional limitation of the current model is the lack of time dynamics, which limits its scope to static one-off predictions.

In future work, we aim to collect a longitudinal dataset and to expand the model to a dynamic BN, in order to enable its application to other clinical reasoning tasks where time is a critical factor, such as prognosis. Further steps will be to explore and document model interpretability, as well as to investigate specificity to MDD by studying control datasets of e.g. bipolar disorder or adjustment disorder.

# 6. References


[1] M. Hunt, J. Auriemma, and A. C. Cashaw, "Self-Report Bias and Underreporting of Depression on the BDI-II," *Journal of Personality Assessment*, vol. 80, no. 1, pp. 26–30, 2003.

[2] S. Vanheule, M. Desmet, H. Groenvynck, Y. Rosseel, and J. Fontaine, "The Factor Structure of the Beck Depression Inventory–II: An Evaluation," *Assessment*, vol. 15, no. 2, pp. 177–187, 2008.

[3] I. S. Marková and G. E. Berrios, "Epistemology of Mental Symptoms," *Psychopathology*, vol. 42, no. 6, pp. 343–349, 2009.

[4] R. Strawbridge, P. McCrone, A. Ulrichsen, R. Zahn, J. Eberhard, D. Wasserman, P. Brambilla, G. Schiena, U. Hegerl, J. Balazs *et al.*, "Care pathways for people with major depressive disorder: A European Brain Council Value of Treatment study," *European Psychiatry*, vol. 65, no. 1, 2022.

[5] J. F. Cohn, N. Cummins, J. Epps, R. Goecke, J. Joshi, and S. Scherer, "Multimodal assessment of depression from behavioral signals," in *The Handbook of Multimodal-Multisensor Interfaces: Signal Processing, Architectures, and Detection of Emotion and Cognition-Volume 2*. Morgan & Claypool, 2018, pp. 375–417.

[6] J. M. Girard and J. F. Cohn, "Automated audiovisual depression analysis," *Current opinion in psychology*, vol. 4, pp. 75–79, 2015.

[7] K. Kroenke, T. W. Strine, R. L. Spitzer, J. B. Williams, J. T. Berry, and A. H. Mokdad, "The PHQ-8 as a measure of current depression in the general population," *Journal of Affective Disorders*, vol. 114, no. 1-3, pp. 163–173, 2009.

[8] W. J. von Eschenbach, "Transparency and the black box problem: Why we do not trust AI," *Philosophy & Technology*, vol. 34, no. 4, pp. 1607–1622, 2021.

[9] S. McLachlan, K. Dube, G. A. Hitman, N. E. Fenton, and E. Kyrimi, "Bayesian networks in healthcare: Distribution by medical condition," *Artificial Intelligence in Medicine*, vol. 107, p. 101912, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0933365720300774

[10] E. Kyrimi, S. McLachlan, K. Dube, M. R. Neves, A. Fahmi, and N. Fenton, "A comprehensive scoping review of bayesian networks in healthcare: Past, present and future," *Artificial Intelligence in Medicine*, vol. 117, p. 102108, 2021.

[11] Z. Yang, H. Li, L. Li, K. Zhang, C. Xiong, and Y. Liu, "Speech-Based Automatic Recognition Technology for Major Depression Disorder," in *The 13th International Conference on Human-Centered Computing*, D. Milošević, Y. Tang, and Q. Zu, Eds. Čačak, Serbia: Springer International Publishing, 2019, pp. 546–553.

[12] C. W. Espinola, J. C. Gomes, J. M. S. Pereira, and W. P. dos Santos, "Detection of major depressive disorder using vocal acoustic analysis and machine learning—an exploratory study," *Research on Biomedical Engineering*, vol. 37, pp. 53–64, 2021.

[13] S. Fara, S. Goria, E. Molimpakis, and N. Cummins, "Speech and the n-Back task as a lens into depression. How combining both may allow us to isolate different core symptoms of depression," in *Proc. Interspeech 2022*. Incheon, Korea: ISCA, 2022, pp. 1911–1915.

[14] M. Valstar, B. Schuller, K. Smith, F. Eyben, B. Jiang, S. Bilakhia, S. Schnieder, R. Cowie, and M. Pantic, "AVEC 2013: The Continuous Audio/Visual Emotion and Depression Recognition Challenge," in *Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge*, ser. AVEC '13. Barcelona, ES: ACM, October 2013, pp. 3–10.

[15] J. Gratch, R. Artstein, G. Lucas, G. Stratou, S. Scherer, A. Nazarian, R. Wood, J. Boberg, D. DeVault, S. Marsella, D. Traum, A. Rizzo, and L.-P. Morency, "The Distress Analysis Interview Corpus of human and computer interviews," in *Proceedings of the 9th International Conference on Language Resources and Evaluation*, ser. LREC '14. Reykjavik, Iceland: ELRA, May 2014, pp. 3123–3128.

[16] S. Nikolin, Y. Y. Tan, A. Schwaab, A. Moffa, C. K. Loo, and D. Martin, "An investigation of working memory deficits in depression using the n-back task: A systematic review and meta-analysis," *Journal of Affective Disorders*, vol. 284, pp. 1–8, 2021.

[17] International Phonetic Association, *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*. Cambridge, UK: Cambridge University Press, 1999.

[18] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan *et al.*, "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2015.

[19] S. Scherer, G. M. Lucas, J. Gratch, A. S. Rizzo, and L.-P. Morency, "Self-reported symptoms of depression and PTSD are associated with reduced vowel space in screening interviews," *IEEE Transactions on Affective Computing*, vol. 7, no. 1, pp. 59–73, 2015.

[20] D. Koller and N. Friedman, *Probabilistic Graphical Models: Principles and Techniques - Adaptive Computation and Machine Learning*. Cambridge, MA, USA: The MIT Press, 2009.

[21] Y. Zeng, S. Shimizu, H. Matsui, and F. Sun, "Causal discovery for linear mixed data," in *Conference on Causal Learning and Reasoning*. Eureka, CA, USA: PMLR, 2022, pp. 994–1009.

[22] W. E. Carson IV, A. Talbot, and D. Carlson, "AugmentedPCA: A Python Package of Supervised and Adversarial Linear Factor Models," https://arxiv.org/abs/2201.02547, 2021.

[23] D. Phan, N. Pradhan, and M. Jankowiak, "Composable Effects for Flexible and Accelerated Probabilistic Programming in NumPyro," https://arxiv.org/abs/1912.11554, 2019.

[24] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[25] V. Berisha, C. Krantsevich, G. Stegmann, S. Hahn, and J. Liss, "Are reported accuracies in the clinical speech machine learning literature overoptimistic?" in *Proc. Interspeech 2022*, 2022, pp. 2453–2457.