Interspeech 2024
1-5 September 2024, Kos, Greece

Revealing Confounding Biases: A Novel Benchmarking Approach for Aggregate-Level Performance Metrics in Health Assessments

Roseline Polle[1], *Salvatore Fara*[1], *Stefano Goria*[1], *Nicholas Cummins*[1,2]

[1]Thymia Limited, London, UK
[2] Institute of Psychiatry, Psychology and Neuroscience, King's College London, UK
roseline@thymia.ai, salvatore@thymia.ai, stefano@thymia.ai, nick.cummins@kcl.ac.uk

Abstract

Numerous speech-based health assessment studies report high accuracy rates for machine learning models which detect conditions such as depression and Alzheimer's disease. There are growing concerns that these reported performances are often overestimated, especially in small-scale cross-sectional studies. Possible causes for this overestimation include overfitting, publication biases and a lack of standard procedures to report findings and testing methodology. Another key source of misrepresentation is the reliance on aggregate-level performance metrics. Speech is a highly variable signal that can be affected by factors including age, sex, and accent, which can easily bias models. We highlight this impact by presenting a simple benchmark model for assessing the extent to which aggregate metrics exaggerate the efficacy of a machine learning model in the presence of confounders. We then demonstrate the usefulness of this model on exemplar speech-health assessment datasets.

Index Terms: Health Assessments, Confound-aware metrics, bias, fairness, computational paralinguistics

1. Introduction

Speech has been identified as a promising biomarker for detecting various clinical conditions, including major depression disorder and Alzheimer's disease [1, 2, 3]. There are numerous recent studies reporting high accuracy rates for speech-based machine learning models as evidence of this claim [4]. However, there are growing concerns that the reported performance of such models is routinely overestimated. Possible causes of this overestimation include overfitting and publication biases [5, 6].

One related concern is the use of aggregate performance metrics [7]. It is widely established in statistics and machine learning theory that aggregate metrics may lead to unexpected conclusions in the presence of confounding variables [8]. A classic example of this phenomenon is Simpson's paradox [9, 10]. This paradox arises when the direction of an association (e. g. the efficacy of a new treatment) between two variables, reverses when data is sliced into different subgroups (e. g. different age groups) [8, 11]. A less explored, but similar confounding effect can be observed in many of the metrics commonly used to assess machine learning results [12, 13].

Such scenarios are potentially concerning in speech-based health assessment research. Speech is a highly variable signal affected by factors such as age, sex, and accent [14]. Moreover, these factors are often strongly linked to the target variables of the models (e.g., depression scores), acting as confounding variables. For example, in large-scale surveys, females often report higher depression rates than males, reaching up to double the prevalence, while younger individuals exhibit higher rates compared to their older counterparts [15].

Taking this confounding argument a step further, it is very much possible that there are specific mechanisms that can yield an over-representation when using aggregate performance metrics. Given the sensitivity of speech, these effects are going to be particularly severe in paralinguistic models. For example, the straightforward scenario where the strength of the target signal is significantly smaller than the strength of the signal coming from confounding variables. We believe that these 'confounder biases' effects on performance metrics contribute to the overestimation of model performance in this space [5].

Understanding the effects of confounder biases is vital in developing models for use in clinical research and practice [16]. To illustrate the impact of such biases on aggregate metrics, we present a simple benchmark model that excels at predicting the confounding variable while failing completely at identifying the target variable. The confound-aware performance scores obtained by these benchmark models – that are higher than what is considered to be chance level – offer a more realistic benchmark for aggregate performance metrics and can be computed for known confounds in simple closed form.

In addressing this, we first highlight on data from [15] that these confound-aware metrics also represent a powerful tool for evaluating reported scores where only demographic information is available. This experiment highlights the importance of our metrics as tools to guide data collection and corpora creation by giving insights on achievable model performance.

We then apply the proposed metric to publicly available *Androids Depression* corpus [17] and the *Pitt Dementia* corpus [18]. To keep the focus on the metrics, we create a simple and standard data pipeline and use publicly available embeddings and simple downstream classification models. Our results provide further arguments against the use of aggregate metrics [7] which align with similar observations in the AI fairness literature [19, 20], but arrive from a different angle and motivation.

2. Methodology

We aim to model a health state from voice recordings by identifying an optimal function ϕ, and parameters θ that express the conditional probability (or density for a regression problem) of a speaker being in a specific health state y as:

$$p(y|a) = \phi(a, \theta) \tag{1}$$

where a denotes embeddings from an audio signal. Now, let's consider a set of M confounding variables $C_0, ..., C_{M-1}$; where we limit the analysis to categorical variables, so each C_i can take values in a finite set of cardinality $n(C_i)$. We can then define the set of all possible values of confounding variables C

10.21437/Interspeech.2024-1092

as the cross product:

$$C_0 \times ... \times C_{M-1} =$$

$$\{(c_0, ..., c_{M-1}) | c_0 \in C_0, ..., c_{M-1} \in C_{M-1}\} \quad (2)$$

Then, if we apply the law of total probability and decompose the conditional distribution across a set of confounding variables $c \in C$, we can re-write the conditional probability (Eq. 1) as:

$$p(y|a) = \sum_{c \in C} p(y|a, c)p(c|a) \quad (3)$$

Given, in many speech-based health assessment scenarios, the mapping of audio embeddings to confounding variables may be an easier task than mapping to the actual target, we can then use Eq. 3 to explore the impact of confounds on a model that (i) accurately captures the relationship between audio embedding and confounding variables, i.e., $p(c|a) = \delta_{c,c_i}$ with δ being a Kronecker delta and c_i is the true confounding variable state for individual i; and (ii) only uses the confounding variable c to predict the target variable, completely disregarding the audio embedding, i.e., $p(y|a, c) = p(y|c)$.

For such a model, the conditional probability for the individual i becomes

$$p(y_i = 1|a) = \sum_{c \in C} p(y_i = 1|c)\delta c, c_i = f_{c_i} \quad (4)$$

where f_{c_i} is the frequency of the health condition class within the demographics group the individual i belongs to.

Note, this is equivalent to a Naïve Bayes model that uses the confounding variables as input.

2.1. Confound-aware classification metrics

Commonly used performance metrics for classification include *Area under the receiver operating characteristic Curve* (AUC), *Balanced Accuracy (BA)* and F_1 *score*. In order to compute those we first need to compute the True and False Positive rates (TPR and FPR) as a function of a decision threshold t. Considering the number of examples as sum of positives and negative $N = N^P + N^N$ we can also write the positives (negatives) as sum of examples in each confounding bucket $N^P = \sum_{c \in C} N_c^P$.

To compute TPR we sum over all the examples in our (test) dataset the model predicted positives and divide by N^P as in

$$TPR(t) = \frac{1}{N^P} \sum_{n=1}^{N} \mathbb{1}_{\{t<p(y_n=1|a)\}} \quad (5)$$

where $\mathbb{1}_{\{t<p(y_n=1|a)\}}$ is 1 if the model probability is higher than the selected threshold.

By grouping all samples that have identical values for the confounding variables, we can restructure the summation across N as follows:

$$\sum_n = \sum_{n \in c_0} + ... + \sum_{n \in c_{M-1}} \quad (6)$$

where with $n \in C_i$ we mean the set of individuals that have the demographics values of C_i. The sum over n in every group becomes trivial:

$$\sum_{n \in c_i} \mathbb{1}_{\{t<p(y_n=1|a)\}} = N_c^P \mathbb{1}_{\{t<f_c\}} \quad (7)$$

and we get that the model defined in the previous section implies

$$\begin{aligned} TPR(t) = & \quad \frac{1}{N^P} \sum_{c \in C} N_c^P \mathbb{1}_{\{t<f_c\}} \\ FPR(t) = & \quad \frac{1}{N^N} \sum_{c \in C} (N_c - N_c^P) \mathbb{1}_{\{t<f_c\}} \end{aligned} \quad (8)$$

From TPR, FPR, N^P and N^N, all relevant performance metrics can be computed as standard. Herein, we will refer to performance metrics built under this idealised model by prepending $c-$ to the metric name as in c-AUC, which stands for *confound-aware AUC*.

Code to compute our metrics will be made available upon acceptance.

2.2. Metrics Interpretation

These confound-aware metrics can differ from their chance-level counterpart; with the size of the impact depending on how well separated the different confound groups are with respect to the health state. Given this, we consider confound-aware metrics as an additional benchmark to be used to put performance into context alongside commonly used benchmarks such as chance-level metrics.

For example, when presenting F_1 scores, it is common to report the F_1 score of a model that randomly selects the target output. Similarly, if reporting BA, it is normal to highlight the accuracy level of a prediction model that simply outputs the majority class. Correspondingly, it is often assumed that AUC should be simply benchmarked against a data independent chance level, such as a 0.5 threshold for a random model. What our confound-aware metrics highlight is that all performance metrics – AUC included – may instead deserve a less forgiving benchmark, especially when dealing with tasks and data with rich confound structure.

In particular, scenarios where confound-aware metrics are considerably above their random counterpart suggest that predictive models may have learnt the wrong signal. Instead of mapping the input to the target variable, such models have learnt a mapping from them the input to confounding variables, that acted as a proxy for the target during training. In those scenarios, the reliance on aggregate performance metrics should be avoided. We recommend in these situations measuring metrics group by group. A well-behaved model will show consistent performance across groups, while a model that indeed only learned the confounded variables effect will show a performance drop at the group level [12, 21].

The opposite scenario is also possible, where confound-aware metrics score lower values than their random counterparts. This scenario suggests that the distribution of confounding variables in a held-out test set is different from the distribution of confounding variables that were used to define the frequencies of the target variables for the various sets c. In this case, if the predictive models we are testing show high aggregate performance metrics, it is very likely that the models are correctly learning the signal. Predictive models performing below chance-level in this scenario suggest overfitting. Equipped with confound-aware metrics, we may be able to be more specific and gain insight into which patterns the model incorrectly picked.

3. Example: Corpus Creation Support

In this example, we explore potential biases within speech models for depression, working under the assumption that the data originates from a representative segment of the US population [15], without adjusting for distinct variables.

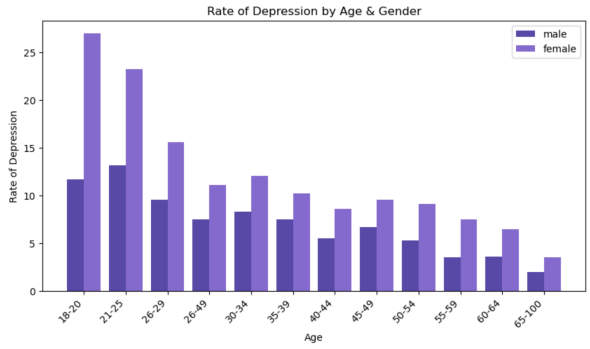

Figure 1: *Distribution of Depression rate in US population*

Depression rates across age and sex groups display quite clear trends (see Fig 1). Following the granularity of the data available in [15] we consider the confounding variables C_0 (sex) and C_1 (age), where age is discretised in 12 groups. The Cartesian product $C = C_0 \times C_1$ is, therefore, a set of 24 elements.

In the application of our proposed metrics for evaluating speech models, we compute the frequencies f_c from the training dataset and ascertain the count of positive and negative instances $N_{P/N}$ using the testing dataset. For this large-scale dataset, encompassing approximately $30k$ individuals, we assume statistical consistency between randomly selected training and testing splits, given the sample's size.

First of all, we compute $c\text{-}AUC = 0.67$, which is significantly higher than chance level 0.5. We also compute $c\text{-}F_1$ and $c\text{-}ACC$ (accuracy) as a function of the decision threshold t (see Fig.2). The values of the c-curves at threshold 0 correspond to the benchmark case "predict only positive" (i.e. minority class here), whereas the value when the curves flatten is "predict only negative". The $c\text{-}F1$ curve peaks at 0.3, above both these benchmarks and chance level, bringing useful additional information. There is no new information for $c\text{-}ACC$ in our case, but there exists scenarios where we could see the same trend as for $F1$.

Employing confound-aware metrics in the corpus creation process, can offer support to develop more balanced and representative speech models by highlighting potential biases within the data even before collecting the recordings.

4. Example: Diagnostic tool for speech models

To illustrate the application of confound-aware metrics as a diagnostic tool for speech-based models, we implement straightforward classification models on public datasets targeting depression and dementia.

4.1. Datasets

We select the *Androids Depression* corpus [17] for depression and the *Pitt Dementia* corpus [18] for dementia; both datasets provide valuable demographics information that allows for investigations on confounding variables effects. Androids consists of two audio recordings from 116 Italian participants. The corpus includes 64 participants previously diagnosed with depression disorder and 52 who had never experienced any mental health issues. All participants conducted an interview and a reading task, we utilised the reading task only in our analysis.

The Pitt Corpus is a dementia study with 291 participants

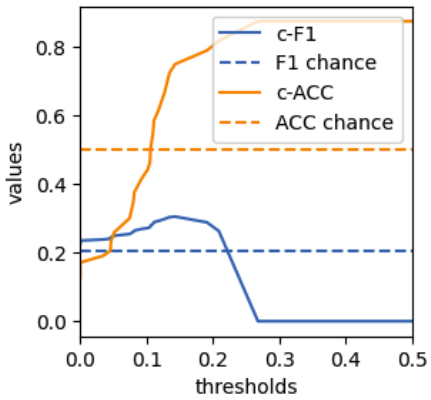

Figure 2: *Comparison of chance level and Confound Aware F_1 ($c\text{-}F_1$) and Accuracy ($c\text{-}ACC$) for Depression Rates in US at different decision thresholds*

Table 1: *Datasets description; confounds are Sex (S), Age (A) and Education (ED)*

Dataset	Target	# participants	Confounds
Androids	Depression	116	S,A,ED
Pitt	Dementia	291	S,A,ED

who completed up to four different allocation tasks. We utilised the *cookie* task for our study, which we split are split into two groups, 98 participants with dementia and 193 matched control.

These datasets are public, all results reported in this paper can be easily reproduced. See Table 1 for a summary of the datasets used.

4.2. Audio preprocessing and modelling

4.2.1. Audio Segmentation

We segmented audio samples into 20-second cuts according to the following steps. transcriptions were generated using Deepgram's Python SDK (whisper-medium model, v3.1.7) and speaker differentiation was achieved via Deepgram's diarization. Participant-specific segments were isolated with timestamps and transcriptions, trimmed to exclude silences over 0.5s, concatenated with 0.1s silences, and then divided into consecutive 20s windows using Lhotse (v1.19.2) [22]. Segments under 20s at file end were discarded.

4.2.2. Embeddings

We derived two embedding types from the audio cuts: the low-dimensional, knowledge-driven eGeMAPS functionals (88 features) via openSMILE [23], and the data-driven TRILLsson4 (T4) embeddings (1024 features) [24]. These were selected for their widespread use and reproducibility, offering complementary insights due to their contrasting dimensionalities and derivation methods.

4.2.3. Classifiers

We used XGB-Classifer (XGBoost v2.0.3) with default parameters as a classifier; we performed 5-fold cross-validation, making sure that there was no overlap of speakers across folds. For the Androids dataset we use the folds that are reported in the benchmark paper [17], while we use random folds for Pitt.

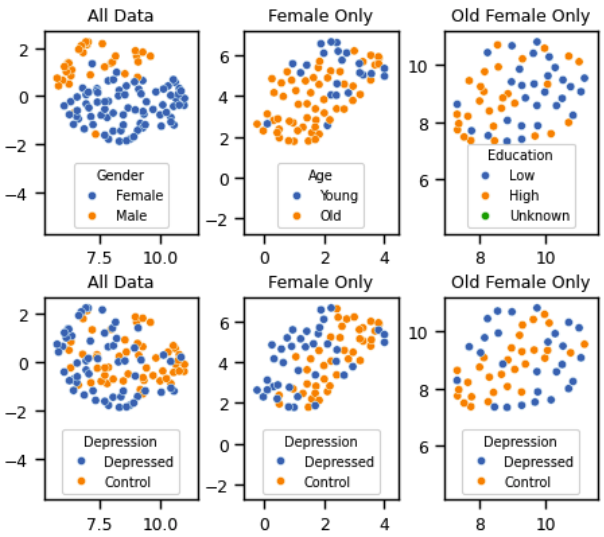

Figure 3: *2D UMAP Visualization of TRILLson4 embeddings in the Androids dataset*

4.3. Embeddings Visualization

To explore the influence of confounding variables versus target variables in our embeddings, we conducted a hierarchical UMAP 2D projection [25], particularly focusing on TRILLson4 embeddings from the Androids dataset. Our findings, illustrated in Fig 3, display binarized confounding variables (top row) and target variables (depression, bottom row). The analysis reveals clear sex separability in the dataset's full 2D UMAP projection. Subsequent columns detail the separation within the female-only subset, showing reduced confounder separability, and then assess the impact of education on the old-female-only subset, where distinguishability between confounding and target variables becomes less apparent. The target variable, depression, appears dispersed across all three dataset splits, indicating that the signal from confounding variables is more pronounced than that of the target variable in this dataset.

4.4. Experiment results

4.4.1. Androids

The confounds available in this corpus are sex (C_0), age (C_1) which we group in *Below 40*, *40-55* and *Over 55*, and education (C_2) on a scale of 1 to 4 which we group in *lower* (1-2) and *higher* (3-4); we also consider $C_{ij} = C_i \times C_j$ as confounding variables.

Table 2 compares our performance results with the two baselines ($BS1$ and $BS2$) from the Androids paper [17]. The confound-aware metrics significantly exceed their random counterparts. Both our benchmark classifiers, and even more notably BS2, exhibit consistently superior performance compared to the confound-aware benchmarks. Although the dataset's small size renders metric performance less reliable when recomputed on data subsets, we still assessed AUC across sex and age groups and found relatively stable performance. This implies that, despite the potential issues posed by confounding effects, the models appear to accurately capture the depression signal.

Table 2: *Androids models and benchmarks comparison. C-Aware metrics act as a new benchmark (see Section 2.2). Note that AUC is not reported in the Androids paper.*

Model	Acc. (%)	F1 (%)	AUC (%)
Rand.	50.1	51.8	-
BS1	69.6 ± 5.3	68.4 ± 7.7	-
BS2	84.4 ± 1.1	83.7 ± 1.1	-
eGeMAPS	73.9 ± 12.6	74.4 ± 13.0	82.1 ± 9.5
TRILLson4	79.8 ± 7.4	80.6 ± 7.8	85.6 ± 8.1
C-Aware	56.1 ± 8.4	58.5 ± 12.2	68.5 ± 10.1

Table 3: *Pitt Dementia; aggregate and age-conditioned AUC compared against c-AUC*

Group	N	AUC (%)
Aggregate	291	74.8 ± 3.7
C-Aware	291	70.2 ± 1.8
Age $<$ 65	115	75.7 ± 14.8
Age 65-74	92	75.7 ± 12.1
Age $>$ 75	84	52.7 ± 29.7

4.4.2. Pitt

The confounds available in the Pitt Corpus are sex (Male and Female), age which we group in *Below 65*, *65-75* and *Over 75*, and education on a scale of 6 to 21 which we group in *Under 13* and *Over 13*; we report results using Age as confound where some biases effects start being visible.

Table 3 shows that aggregate AUC $= 0.748$ for our model targeting dementia and compares it against c-AUC $= 0.702$ and AUC obtained on the three age splits we used as confounding variables. We see in this case that the c-AUC value is quite close to the aggregate AUC from the speech model and the performance conditioned on the *Over 75* age bucket approaches chance level. In our idealised counfound-aware model described in Section 2, AUC for the 3 age groups would be 0.5. While the deviation observed here is less extreme, it still indicates a potential model failure mode.

5. Conclusion

In our study, we demonstrated the application of confound-aware metrics through two distinct use cases: aiding in the creation of unbiased speech corpora and serving as a diagnostic tool for evaluating speech models targeting depression and dementia. Our findings underscore the prominence of confounding signals over target variables and the limitations of relying solely on aggregate metrics for evaluation, especially in health applications. Speech data can inadvertently capture common confounds, suggesting that a multi-dimensional evaluation framework is necessary to prevent the overestimation of model performances [5, 6]. However, further research with larger datasets would be beneficial to more thoroughly examine the extent to which speech models are influenced by the potential issues indicated by the metrics.

Our research emphasizes the need for a holistic approach, integrating confound-aware metrics with conventional methods, to provide a thorough and accurate assessment of speech-based health models. There is an urgent need to integrate such metrics into reporting standards. Such next steps are crucial in developing more effective and unbiased diagnostic tools in the healthcare domain [16].

6. Acknowledgements

NC: This paper also represents independent research part funded by the National Institute for Health Research (NIHR) Maudsley Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London and supported by the National Institute for Health and Care Research University College London Hospitals Biomedical Research Centre. Views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Dept. of Health and Social Care.

7. References

[1] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. F. Quatieri, "A review of depression and suicide risk assessment using speech analysis," *Speech communication*, vol. 71, pp. 10–49, 2015. [Online]. Available: https://doi.org/10.1016/j.specom.2015.03.004

[2] D. M. Low, K. H. Bentley, and S. S. Ghosh, "Automated assessment of psychiatric disorders using speech: A systematic review," *Laryngoscope investigative otolaryngology*, vol. 5, no. 1, pp. 96–116, 2020. [Online]. Available: https://doi.org/10.1002%2Flio2.354

[3] S. De la Fuente Garcia, C. W. Ritchie, and S. Luz, "Artificial intelligence, speech, and language processing approaches to monitoring alzheimer's disease: a systematic review," *Journal of Alzheimer's Disease*, vol. 78, no. 4, pp. 1547–1574, 2020. [Online]. Available: https://doi.org/10.1049/cit2.12113

[4] P. Wu, R. Wang, H. Lin, F. Zhang, J. Tu, and M. Sun, "Automatic depression recognition by intelligent speech signal processing: A systematic survey," *CAAI Transactions on Intelligence Technology*, vol. 8, no. 3, pp. 701–711, 2023. [Online]. Available: https://doi.org/10.1049/cit2.12113

[5] 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*. Incheon, Korea: ISCA, 2022, pp. 2453–2457. [Online]. Available: https://doi.org/10.21437/Interspeech.2022-691

[6] M. Fire and C. Guestrin, "Over-optimization of academic publishing metrics: observing goodhart's law in action," *GigaScience*, vol. 8, no. 6, p. giz053, 2019. [Online]. Available: https://doi.org/10.1093/gigascience/giz053

[7] R. Burnell, W. Schellaert, J. Burden, T. D. Ullman, F. Martinez-Plumed, J. B. Tenenbaum, D. Rutar, L. G. Cheke, J. Sohl-Dickstein, M. Mitchell, D. Kiela, M. Shanahan, E. M. Voorhees, A. G. Cohn, J. Z. Leibo, and J. Hernandez-Orallo, "Rethink reporting of evaluation results in AI," *Science*, vol. 380, no. 6641, pp. 136–138, 2023. [Online]. Available: https://doi.org/10.1126/science.adf6369

[8] J. Pearl, M. Glymour, and N. P. Jewell, *Causal inference in statistics: a primer*. Chichester, West Sussex: Wiley, 2016, ch. 3.

[9] S. Ameringer, R. C. Serlin, and S. Ward, "Simpson's paradox and experimental research," *Nursing research*, vol. 58, no. 2, pp. 123–127, 2009. [Online]. Available: https://doi.org/10.1097%2FNNR.0b013e318199b517

[10] S. A. Julious and M. A. Mullee, "Confounding and simpson's paradox," *The BMJ*, vol. 309, no. 6967, pp. 1480–1481, 1994. [Online]. Available: https://doi.org/10.1136/bmj.309.6967.1480

[11] M. L. Samuels, "Simpson's paradox and related phenomena," *Journal of the American Statistical Association*, vol. 88, no. 421, pp. 81–88, 1993. [Online]. Available: https://doi.org/10.2307/2290700

[12] L. Dixon, J. Li, J. Sorensen, N. Thain, and L. Vasserman, "Measuring and Mitigating Unintended Bias in Text Classification," in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, 2018, p. 67–73. [Online]. Available: https://doi.org/10.1145/3278721.3278729

[13] R. Sharma, H. Garayev, M. Kaushik, S. A. Peious, P. Tiwari, and D. Draheim, "Detecting Simpson's Paradox: A Machine Learning Perspective," in *International Conference on Database and Expert Systems Applications*. Springer, 2022, pp. 323–335. [Online]. Available: https://doi.org/10.1007/978-3-031-12423-5_25

[14] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. MüLler, and S. Narayanan, "Paralinguistics in speech and language—state-of-the-art and the challenge," *Computer Speech & Language*, vol. 27, no. 1, pp. 4–39, 2013. [Online]. Available: https://doi.org/10.1016/j.csl.2012.02.005

[15] National Institute of Mental Health, "Prevalence of major depressive episode among adults," 2020. [Online]. Available: https://www.nimh.nih.gov/health/statistics/major-depression

[16] M. E. McNamara, M. Zisser, C. G. Beevers, and J. Shumake, "Not just "big" data: Importance of sample size, measurement error, and uninformative predictors for developing prognostic models for digital interventions," *Behaviour research and therapy*, vol. 153, p. 104086, 2022. [Online]. Available: https://doi.org/10.1016/j.brat.2022.104086

[17] F. Tao, A. Esposito, and A. Vinciarelli, "The Androids Corpus: A New Publicly Available Benchmark for Speech Based Depression Detection," in *Proc. INTERSPEECH 2023*. Dublin, Ireland: ISCA, 2023, pp. 4149–4153. [Online]. Available: https://doi.org/10.21437/Interspeech.2023-894

[18] J. T. Becker, F. Boiler, O. L. Lopez, J. Saxton, and K. L. McGonigle, "The natural history of alzheimer's disease: description of study cohort and accuracy of diagnosis," *Archives of Neurology*, vol. 51, no. 6, pp. 585–594, 1994. [Online]. Available: https://doi.org/10.1001/archneur.1994.00540180063015

[19] N. Kallus and A. Zhou, "The Fairness of Risk Scores Beyond Classification: Bipartite Ranking and the xAUC Metric," https://arxiv.org/abs/2305.15614, 2019.

[20] H. Fong, V. Kumar, A. Mehrotra, and N. K. Vishnoi, "Fairness for AUC via Feature Augmentation," https://arxiv.org/abs/2111.12823.

[21] D. Borkan, L. Dixon, J. Li, J. Sorensen, N. Thain, and L. Vasserman, "Limitations of Pinned AUC for Measuring Unintended Bias," https://arxiv.org/abs/1903.02088, 2019.

[22] P. Żelasko, D. Povey, J. Trmal, S. Khudanpur *et al.*, "Lhotse: a speech data representation library for the modern deep learning ecosystem," https://arxiv.org/abs/2110.12561, 2021.

[23] F. Eyben, M. Wöllmer, and B. Schuller, "The Munich Versatile and Fast Open-Source Audio Feature Extractor," in *Proceedings of the 18th ACM international conference on Multimedia*. Melbourne , VIC , Australia: ACM, 2010, pp. 1459–1462. [Online]. Available: https://doi.org/10.1145/1873951.1874246

[24] J. Shor and S. Venugopalan, "Trillsson: Distilled universal paralinguistic speech representations," https://arxiv.org/abs/2203.00236, 2022.

[25] I. Ben-Shaul, R. Shwartz-Ziv, T. Galanti, S. Dekel, and Y. LeCun, "Reverse Engineering Self-Supervised Learning," https://arxiv.org/abs/2305.15614, 2023.