Title

The Illusion of Detection: Why Machine Learning Can Overestimate Mental Disorder Recognition from Audio

Authors

Sia Sha, Pamela Qualter, Yongchao Huang, Nina Shahrizad, Dario Krpan, Matteo Galizzi

Abstract

Machine learning (ML) research has attempted to detect mental disorders from people's audio recordings. Despite seemingly impressive reported results, it remains unclear whether those results reflect genuine clinical signal or artefacts related to how ML models are trained and evaluated. We re-examine 25 studies included in a recent systematic review and identify three major recurring methodological problems that are likely to inflate the ability of ML models to detect mental disorders: (1) data leakage, (2) lack of multiple seeds, and (3) lack of large, naturalistic datasets. We show via an experiment how ML practices can produce apparently strong results even in the absence of real, predictive signals. We also conduct robust analyses using a large, novel dataset, where we find no reliable evidence that mental disorders can be detected from audio across multiple modelling approaches. Our findings suggest that current evidence may substantially overestimate the real-world capability of audio-based detection of mental health disorders, highlighting the need for evaluation standards aligned with best practices for clinical and policy use.

Introduction

Roughly one billion people suffer from mental disorders [1]. Measurement of these disorders is currently done via clinical interviews or self-report questionnaires [2]. While effective, those approaches require trained professionals, time, and sustained patient engagement, making large scale screening difficult [3], [4]. As a result, researchers have increasingly explored whether machine learning (ML) could assist detection of mental health disorders using passively collected data. Researchers, particularly, have focused on training ML models on speech, since it is easy to collect, and clinicians have long believed that mood and mental health may influence vocal characteristics [5]. For example, the speech of patients with major depressive disorder is often described as monotone, delayed, and slow [6].

Recent ML studies report high classification accuracy [7], suggesting that mental disorders could be detected automatically from voice recordings, and raising the possibility of scalable and low-burden mental health assessment. If reliable, such systems could have substantial implications for clinical practice and public health policy, including earlier identification of risk and improved access to support. However, the practical value of automated detection depends not only on performance within research datasets, but on whether models generalise to new individuals across diverse, real-world environments. A central challenge in ML research is that reported accuracy depends heavily on how models are trained and evaluated. Certain methodological practices can unintentionally allow algorithms to exploit patterns unrelated to mental health itself, producing optimistic performance estimates that do not translate beyond controlled datasets [8]. For clinicians and policymakers, this distinction corresponds to the difference between apparent accuracy under controlled research conditions and reliable performance in practice.

Despite rapid growth in audio-based mental health research, it remains unclear whether reported success reflects genuine predictive signal or artefacts related to how models are trained and evaluated. This paper directly tests the assumption that mental disorders can be reliably inferred from general speech recordings under realistic conditions. We make two contributions. First, we conduct a critical re-evaluation of 25 studies included in a recent systematic review focused on depression [9], and we identify substantial and previously undocumented flaws. We also provide an experiment to illustrate recurring methodological practices systematically inflate reported performance. Second, we evaluate whether reliable detection remains possible when those issues are removed. Using a large, diverse dataset collected across naturalistic settings,

we implement corrected and robust analyses across multiple modelling approaches. Together, those analyses provide a direct test of whether current evidence reflects genuine detectability or flaws in model training.

Evaluation Standards

Throughout this study, we define successful detection as performance that generalises to previously unseen individuals, ideally recorded under heterogeneous, real-world conditions. Such conditions include variation in recording devices, environments, and participant characteristics, and do not rely on structured clinical prompting or dataset-specific artefacts. This definition reflects the conditions under which automated detection systems would ultimately need to operate in real-world clinical and policy settings.

Findings

Evaluation of 25 studies included in a recent systematic review

We re-analyzed 25 studies included in a recent systematic review focused on depression [9]. Across studies, we identified recurring themes: 1) data leakage, 2) lack of multiple seeds, and 3) robustness of results under naturalistic conditions.

General data

Table 1 provides general data for the 25 studies we re-evaluated [10]–[34]. A key takeaway is that 20 studies (80%) used the same dataset, called the "DAIC-WOZ" (Distress Analysis Interview Corpus - Wizard of Oz). This means that potential idiosyncrasies of this dataset—rather than genuine, generalisable depression signals—may be amplified across the literature. The DAIC-WOZ, moreover, is a small dataset: it contains 189 one-to-one clinical interviews between an animated artificial intelligence (AI) agent controlled by a human in one room, and a participant in another room [35]. Importantly, participants in the DAIC-WOZ dataset are not just from the general population; instead, a substantial number of participants are army veterans who likely exhibit particularly poor mental health [35].

Table 1: General information

Study	**Year**	**Dataset**
Bhavya et al	2023	DAIC-WOZ
Chlasta et al	2019	DAIC-WOZ
Cui et al	2022	DAIC-WOZ
Das et al	2024	DAIC-WOZ, MODMA, RAVDESS
Du et al	2023	DAIC-WOZ, MODMA
Feng et al	2023	DAIC-WOZ
Gupta et al	2023	SH2-FS, DAIC-WOZ
Homsiang et al	2022	DAIC-WOZ
Ishimaru et al	2023	DAIC-WOZ
Jenei et al	2020	HDSD
Jenei et al	2021	HDSD
Marriwala et al	2023	DAIC-WOZ
Othmani et al	2021	DAIC-WOZ, RECOLA
Ravi et al	2022	DAIC-WOZ, CONVERGE
Ravi et al	2024	DAIC-WOZ, EATD
Rejaibi et al	2022	DAIC-WOZ, RAVDESS, AVi-D
Saidi et al	2020	DAIC-WOZ
Sardari et al	2022	DAIC-WOZ
Suparatpinyo et al	2023	Thai Corpus
Tian et al	2023	DAIC-WOZ

Study	Year	Dataset
Wang et al	2022a	CHD
Wang et al	2022b	Custom dataset
Yang et al	2023	NRAC, DAIC-WOZ
Yin et al	2023	DAIC-WOZ, MODMA
Zhou et al	2022	DAIC-WOZ

Performance at face value

Table 2 shows how well mental health can be predicted from audio if the literature is taken at face value. In Table 2, we report accuracy, sensitivity, and specificity, which we directly extracted from [9]. Results for accuracy, sensitivity, and specificity appear to mirror and exceed those of established depression detection tools, such as the nine-item patient health questionnaire (PHQ-9) [36].

However, because these measures are sensitive to class imbalances and may therefore be overly optimistic, we also computed Youden's J. Youden's J ranges from -1 to 1. A Youden's J value of 1 indicates that a model does not yield any false positives or false negatives, a value of 0 indicates that a model is essentially guessing randomly, and a value of less than 0 indicates that a model does worse than random guessing. Youden's J reveals a more muted performance for the current literature, although the performance still "seems" acceptable.

Table 2: Performances

Study	Year	Accuracy	Sensitivity	Specificity	Youden's J
Bhavya et al	2023	0.7	0.79	0.5	0.29
Chlasta et al	2019	0.78	0.14	1.0	0.14
Cui et al	2022	0.91	0.82	0.94	0.76
Das et al	2024	0.9	0.94	0.86	0.80
Du et al	2023	0.86	0.9	0.82	0.72
Feng et al	2023	0.77	0.58	0.87	0.45
Gupta et al	2023	0.99	0.99	0.96	0.95
Homsiang et al	2022	0.95	1.0	0.9	0.90
Ishimaru et al	2023	0.97	0.95	0.98	0.93
Jenei et al	2020	0.84	0.8	0.88	0.68
Jenei et al	2021	0.85	0.88	0.83	0.71
Marriwala et al	2023	0.57	0.63	0.42	0.05
Othmani et al	2021	0.73	0.46	0.84	0.30
Ravi et al	2022	0.74	0.77	0.67	0.44
Ravi et al	2024	0.83	0.67	0.91	0.58
Rejaibi et al	2022	0.54	0.35	0.59	-0.06
Saidi et al	2020	0.68	0.71	0.65	0.36
Sardari et al	2022	0.7	0.68	0.72	0.40
Suparatpinyo et al	2023	0.77	0.74	0.8	0.54
Tian et al	2023	0.88	0.86	0.88	0.74
Wang et al	2022a	0.84	0.88	0.81	0.69
Wang et al	2022b	0.75	0.76	0.74	0.50
Yang et al	2023	0.87	0.91	0.76	0.67
Yin et al	2023	0.94	0.92	0.95	0.87
Zhou et al	2022	0.83	0.83	0.83	0.66

Critical evaluation

Next, we critically re-evaluated the 25 studies included in the systematic review—by auditing their reported methodology and, where available, their published code—in relation to three issues: 1) data leakage, 2) lack of multiple seeds, and 3) robustness of results by evaluating under naturalistic conditions.

Issue 1: Data leakage

In ML, models are trained on an initial chunk of data, called the "train set". Afterwards, models are evaluated on another chunk of data, called the "test set". This test set is supposed to be "independent" and "unseen", meaning that no part of it sneaks into the train data. When, however, the train set is contaminated with data from the test set, this is called "data leakage", which is a problem because it inflates results [37].

It turns out that of the 25 studies included in the systematic review, 12 (48%) suffered from data leakage or were at high risk of leakage. Generally, data leakage occurred because models were evaluated on the same speakers that they were trained on. This is akin to giving pupils the same questions during lessons and when sitting exams. Furthermore, studies regularly chopped a speaker's transcript into multiple segments and then distributed these across train and test sets, which means that the same speaker could leak more than once, exacerbating inflated performance. Table 3 shows the exact reasons why speaker leakage occurred or why it was a real risk in each study.

Table 3: Investigation of Data Leakage

Study	**Year**	**Leakage source**
Bhavya et al	2023	No evidence
Chlasta et al	2019	Validation/test set subset of train set.
Cui et al	2022	Audio cut into 15s segments and likely distributed across sets.
Das et al	2024	Train and validation loss curves match suspiciously.
Du et al	2023	MODMA participants distributed over train and validation/test set(s).
Feng et al	2023	No evidence
Gupta et al	2023	No evidence
Homsiang et al	2022	Same participants distributed over train and validation/test set(s).
Ishimaru et al	2023	Setting 1 distributes the same people over train and test sets.
Jenei et al	2020	No evidence
Jenei et al	2021	No evidence
Marriwala et al	2023	No evidence
Othmani et al	2021	Code distributes same speakers across train and validation/test sets.
Ravi et al	2022	Entire paper focuses on identifying speakers.
Ravi et al	2024	No evidence
Rejaibi et al	2022	Train and validation loss curves match suspiciously.
Saidi et al	2020	No evidence
Sardari et al	2022	No evidence
Suparatpinyo et al	2023	Same participants distributed over train and validation/test set(s).
Tian et al	2023	No evidence
Wang et al	2022a	No evidence
Wang et al	2022b	Same participants distributed over train and validation/test set(s).
Yang et al	2023	No evidence
Yin et al	2023	Same participants distributed over train and validation/test set(s).
Zhou et al	2022	No evidence

To demonstrate how speaker leakage can inflate performance, we designed an experiment using a k-nearest neighbours (KNN) classifier, which predicts a data point's label based on its similarity to previously seen data points. We chose KNN because it isolates the underlying mechanism directly: any performance gain when a speaker appears in both train and test sets can be attributed specifically to the model matching that

speaker's voice to itself, rather than to any genuine depression signal. Our experiment, described in detail in the Methods section, had two conditions, averaged across 50 runs. In the first condition, any individual speaker could either appear in the train set or the test set, but not both. In the second condition, speakers appeared in both train and test sets. When individual speakers appeared in both sets—that is, they leaked across sets—the performance for accuracy was somewhat inflated, and the performance for sensitivity and Youden's J was substantially inflated. Table 4 reports our results.

Table 4: KNN classifier performance with and without speaker leakage between train and test sets, averaged across 50 runs

Condition	**Accuracy**	**Sensitivity**	**Specificity**	**Youden's J**
KNN	0.65	0.42	0.68	0.11
KNN with leakage	0.67	0.60	0.68	0.28

But why does speaker leakage cause inflated performance? Our response is that it appears that models memorise speaker identity, associating certain speakers with a certain outcome, rather than recognising genuine depression patterns [38].

Issue 2: Lack of multiple seeds

ML models rely on a random number generator to make their predictions, and the sequence of random numbers that this generator produces is determined by a "seed". This seed is the initial state of the random number generator, and to set this initial state, researchers assign some integer value such as "seed = 42" in their research code. Although the exact value of this seed seems benign, the performance of models can vary substantially depending on the chosen seed. While "seed = 42" could yield good results, "seed = 32" could result in a useless model [39]. For example, in our above KNN experiment across 50 seeds, the worst performing seed achieved an accuracy of 0.50, which is essentially random performance. The best performing seed, on the other hand, achieved an accuracy of 0.76.

Robust ML research, therefore, tests the same model multiple times varying the seeds, and then averages the performance of the model across the seeds [40]. Nevertheless, only one of the 25 re-reviewed studies made use of multiple seeds [32]. Reported results in the literature, therefore, could simply be due to particularly favourable seeds. A single seed introduces the risk of "seed-hacking", i.e. the practice of trying various seeds, until one yields desired results [39].

Issue 3: Robustness of results under naturalistic conditions

To test whether reliable detection persists when potential training flaws are removed, we conducted robust analyses using the Bridge2AI dataset. The Bridge2AI dataset contains overall 833 participants recruited across specialty clinics in the US [41]. The researchers measured outcomes such as depression using standardised scales such as the PHQ-9, and they recorded vocal data in heterogeneous real-world recording environments [41]. Using common cut-off points, Table 5 shows the binary distribution for participants in the Bridge2AI dataset who are positive and negative across mental health outcomes. The table shows raw counts and percentages in parentheses. Since not every person participated in the mental health arm of the research, the counts of participants are less per outcome than 833.

Table 5: Prevalence of depression, anxiety, and loneliness in the Bridge2AI sample

Outcome	**Count**	**Positive**	**Negative**
Depression	773	100 (13%)	678 (87%)
Anxiety	774	86 (11%)	688 (89%)
Loneliness	163	56 (34%)	107 (66%)

We used three types of analyses to explore the robustness of mental health detection from audio in the Bridge2AI dataset. We used correlational analysis, static classifiers, and time-series modelling. Our approaches mirrored those in the studies we reviewed. Overall, we found no reliable evidence that acoustic features predicted mental disorders.

Table 6 shows the Spearman's correlation coefficients between mental health outcomes and eleven variables that appeared promising in the literature [42], [43]. For example, "speaking_rate" is the number of words spoken in the recording, and "F0semitoneFrom27.5Hz_sma3nz_stddevNorm" is a measure of pitch variability. The reader will find descriptions of other variables in the literature [44]. Almost all Spearman's correlations in Table 6 are close to 0, indicating no monotonic relationships between variables and outcomes. Two variables, to be fair, did appear to weakly correlate with loneliness. Rather than these being genuine correlations, however, we considered the coefficients to be artefacts of a small sample size.

Table 6: Spearman's correlation coefficients between acoustic variables and mental health outcomes in the Bridge2AI dataset

Variable	**Depression**	**Anxiety**	**Loneliness**
mfcc1_sma3_stddevNorm	0.07	0.06	-0.05
mfcc2_sma3_stddevNorm	-0.06	-0.02	0.02
mfcc3_sma3_stddevNorm	0.02	0.08	-0.05
mfcc4_sma3_stddevNorm	0.03	-0.00	-0.10
F1frequency_sma3nz_amean	-0.01	0.02	0.11
F2frequency_sma3nz_amean	-0.00	0.02	0.07
F0semitoneFrom27.5Hz_sma3nz_stddevNorm	-0.05	-0.04	0.09
cepstral_peak_prominence_mean	0.00	-0.00	0.03
local_jitter	-0.02	-0.00	-0.06
mean_hnr_db	0.04	0.04	0.06
speaking_rate	-0.08	-0.03	-0.05

Table 7 shows the performance of static classifiers for mental health outcomes on the Bridge2AI data [10]. The table shows mean performance across seeds with standard deviations in parentheses. We selected static classifiers that seemed to have done well in previous research, and we used the eleven variables from Table 6 as inputs. Although accuracies in Table 7 might suggest models can meaningfully classify outcomes, Youden's J, which is insensitive to class imbalance, suggests essentially random performance. Interestingly, the random forest model, the classifier with the highest accuracy, appeared to default to predicting the majority class.

Table 7: Static classifier performance for depression, anxiety, and loneliness in the Bridge2AI dataset. Values show mean (SD) across seeds.

Model	**Metric**	**Depression**	**Anxiety**	**Loneliness**
Logistic regression	Accuracy	0.57 (0.04)	0.58 (0.04)	0.51 (0.08)
	Sensitivity	0.49 (0.10)	0.41 (0.11)	0.40 (0.11)
	Specificity	0.59 (0.05)	0.60 (0.05)	0.57 (0.11)
	Youden's J	0.08 (0.10)	0.01 (0.11)	-0.03 (0.15)
Random forest	Accuracy	0.78 (0.03)	0.83 (0.03)	0.54 (0.07)
	Sensitivity	0.14 (0.07)	0.14 (0.09)	0.36 (0.13)
	Specificity	0.88 (0.03)	0.91 (0.03)	0.64 (0.10)
	Youden's J	0.01 (0.06)	0.05 (0.09)	-0.00 (0.15)
Support vector machine	Accuracy	0.66 (0.04)	0.68 (0.04)	0.52 (0.08)
	Sensitivity	0.31 (0.10)	0.28 (0.09)	0.29 (0.14)
	Specificity	0.71 (0.05)	0.73 (0.05)	0.64 (0.12)
	Youden's J	0.02 (0.10)	0.01 (0.09)	-0.07 (0.16)

Table 8 shows the performance of our time-series modelling on unseen test data after ten epochs of training. The time-series models used acoustic properties from a closed speech task to attempt the prediction of mental health. The Table 8 shows similar results to Table 7: some apparent discriminant behaviour based on accuracy, but essentially random or worse-than-random predictions once Youden's J is considered. Notably, performance on the train data was improving with every epoch of training. On the loneliness data, for example, sensitivity, specificity, and Youden's J improved from 0.62, 0.35, and -0.03 in epoch one to 0.89, 0.9, 0.79 respectively in epoch ten. The models, however, over-fitted on unseen test data: for example, while models achieve Youden's J scores of -0.03, 0.17, 0.34 in epochs one, two, and three for training data, the models achieved scores of -0.04, -0.06, and -0.09 on test data in the same epochs respectively for loneliness. All this suggests that our models may have learned to recognise speakers, rather than learn generalisable acoustic markers of mental health status. On unseen data, the models often appeared to default to majority class prediction.

Table 8: Time-series classifier performance for depression, anxiety, and loneliness in the Bridge2AI dataset. Values show mean (SD) across seeds.

Metric	**Depression**	**Anxiety**	**Loneliness**
Accuracy	0.77 (0.04)	0.80 (0.04)	0.51 (0.07)
Sensitivity	0.14 (0.11)	0.16 (0.09)	0.24 (0.14)
Specificity	0.86 (0.06)	0.88 (0.05)	0.65 (0.10)
Youden's J	-0.00 (0.08)	0.03 (0.06)	-0.11 (0.16)

Discussion

This study revisits a rapidly expanding literature suggesting that mental disorders can be detected automatically from speech. Our findings indicate that much of the reported success may arise from the choice of performance metrics as well as the way ML models are trained. Performance metrics are sensitive to class imbalances, and usually mental health datasets contain few positive cases of participants with mental health disorders. On the other hand, the lack of multiple seeds to evaluate models and data leakage are likely inflating results some studies report. In contrast, when we evaluated the classification of mental health status under robust experimental settings, we did not observe reliable mental health detection from acoustic markers in our data.

Our findings, however, do not imply that speech necessarily lacks practical relevance. Rather, the results suggest that current ML approaches need to be aligned with robust evaluation standards to make meaningful progress for the automated detection of mental disorders at population scale. While the field of traditional statistics has by now developed an understanding of its methodological pitfalls, e.g. small sample sizes inflating results [45], modern ML is a much younger discipline, and yet needs to develop its own understanding of pitfalls [46].

Our findings have important implications for clinical practice and policy. Premature deployment of unreliable automated screening tools could lead to false reassurance, unnecessary referrals, misallocation of limited resources, and erosion of trust in digital mental health technologies [47]. Therefore, careful evaluation standards are essential before integrating such systems into healthcare pathways. More broadly, the challenges identified here extend beyond audio-based detection. Many areas of AI in healthcare rely on benchmark datasets and flexible modelling pipelines that may unintentionally favour optimistic conclusions [48]. Existing evidence-assessment frameworks were developed for traditional statistical studies and often overlook ML-specific risks such as data leakage. As a result, systematic reviews may aggregate inflated findings without detecting shared methodological weaknesses. Progress in this area will likely depend on evaluation standards aligned with deployment conditions, including participant-level separation, heterogeneous data collection, repeated stochastic evaluation, and transparent reporting of model selection procedures. Establishing such standards is not only a technical issue, but also a governance and policy priority.

Conclusion

The promise of scalable mental health assessment remains compelling. However, scientific progress depends on distinguishing genuine signal from artefact. Our re-evaluation found that 12 of the 25 reviewed studies (48%) showed evidence of, or were at high risk of, data leakage, and only one study used multiple seeds to guard against seed-hacking. When we corrected for these issues and evaluated detection under naturalistic conditions using a large, independent dataset, we found no reliable evidence that mental disorders can be detected from audio. Our findings suggest that current claims regarding audio-based detection should be interpreted cautiously, and that stronger methodological standards are necessary before such systems can be considered reliable for real-world use. We recommend that future work distribute different participants across train-test splits to eliminate speaker leakage, rather than chop a participant's recording into multiple segments and distribute these across sets. We also recommend that studies report model performance averaged across multiple seeds rather than a single run, that studies use evaluation metrics such as Youden's J that are insensitive to class imbalances, and that studies employ large, naturalistic datasets.

Declarations

Responsibilities

SS, PQ, and NS wrote the draft. YH, DK, and MG advised on the methodology and edited the draft. SS wrote the code.

Funding

None

Code availability

Our code is available here.

Method

Study selection for review

We included all 25 studies from a recent systematic review for re-evaluation [9]. Detailed exclusion and inclusion criteria can be found in the original review manuscript [9]. We chose this particular review, because it had filtered for better-quality studies with more transparent reporting. We then re-evaluated the included studies by parsing their manuscripts and, where available, their published code bases.

Performance metric calculations

We present the face-value performance of classifiers in our re-examined studies, using values as reported in the systematic review [9]. In addition, we calculated Youden's J, a performance metric insensitive to class imbalance [49].

$$J = \text{sensitivity} + \text{specificity} - 1 \tag{1}$$

We report Youden's J because the datasets used in the literature are heavily imbalanced, with relatively few positive cases for the disorders studied. This imbalance can make metrics such as accuracy overly optimistic estimates of performance. We report the same metrics for our own modelling.

Effect of using different seeds

We used 50 seeds for our modelling, except for the time-series models where we used 20 seeds to reduce compute time. We chose a floor of 20 seeds to ensure sufficient statistical power for reliable estimation of mean performance [50]. Both model initialisation and data splits varied across seeds. When reporting results, we averaged the performance of our models across seeds.

Bridge2AI dataset

The Bridge2AI dataset is an ongoing project [51]. The project seeks to create an ethically sourced flagship dataset for future ML research into voice as a biomarker. Participants had presented at five speciality clinics in North America during routine screening or treatment. They became eligible for participation based on known disorders, such as neurological, affective, or vocal [51]. Once participants consented to the study, researchers collected standardised data on demographics, health behaviour, and various vocal measurements. In its current form, Version 3.0, the dataset includes a total of 833 participants, although the exact number of participants for our analyses varies across outcomes.

Outcomes

We used three mental health outcomes: depression, anxiety, and loneliness. Participants on the Bridge2AI dataset completed the PHQ-9 as a measure of depression, the seven-item generalized anxiety disorder questionnaire as a measure of anxiety, and a custom single-item measure for loneliness [51]. We binarized variables such that each outcome had a positive ($y = 1$) and a negative ($y = 0$) class.

$$y \in \{0, 1\} \tag{2}$$

We used common cut-off points from previous work or our own conservative cut-offs to classify participants based on their raw questionnaire scores [52].

$$y = \begin{cases} 1 & \text{if } s \geq \tau \\ 0 & \text{if } s < \tau \end{cases}, \tag{3}$$

where τ is a cut-off point specific to the outcome. Our cut-off points were $\tau_{\text{depression}} \geq 10$, $\tau_{\text{anxiety}} \geq 10$, and $\tau_{\text{loneliness}} \geq 5$.

Predictors for KNN experiment and time-series models

We used Mel frequency cepstrum coefficients (MFCCs) as predictors for our KNN experiment and time-series models, because the literature, as it currently stands, believes that they are highly predictive [9]. In practical terms, MFCCs are inputs that measure the presence of certain speech frequencies at any given time in a recording. These MFCCs were computed from "spectrograms", which themselves are a visual representation of frequencies in raw audio files. The raw audio files that were used to create spectrograms, in turn, came from two tasks: a Rainbow task, and a story-recall speech task. Data from open speech tasks was not available [51]. In the Rainbow task, participants were asked to read a few sentences about how light reflects and how, according to folklore, there is a pot of gold at the end of rainbows [53]. In the story-recall task, participants listened to a story and then recalled its details. The Rainbow task was more constrained, and the story recall task was more unconstrained. Since results were essentially the same, we only report results for the Rainbow task in the manuscript. Initially, we had decided to use data from both tasks, because more unrestricted data may be more predictive of outcomes, although more restricted tasks may be more feasible in deployment settings [54].It is important to note that recording conditions varied, partially due to the different clinical sites themselves, but also because different recording and input devices were used, e.g. with and without headset [51].

We did not compute these MFCCs ourselves, as the raw data was unavailable to us. Instead, the original researchers precomputed and released MFCCs using the Torchaudio interface [51]. On a technical level, these MFCCs were derived from spectrograms computed using a short-time fast Fourier transform with a 25ms window size and 10ms hop length [55]. Mathematically, MFCCs were input matrices for our models $\mathbf{X} \in \mathbb{R}^{T \times 60}$ with 60 dimensions per time frame t with $t = 1, \cdots, T$. The exact value of T varied from sample to sample due to recording length.

We normalised the MFCC vectors. In some instances, it was cosmetic. In other instances, modelling would have failed without it. To normalise, we calculated means and standard deviations for each component of the MFCC vector across all timesteps T. These calculations were based on the train set only—otherwise, there would have been data leakage—and then we used these summary statistics to normalise all datasets.

$$\bar{\mathbf{x}} = \mathbb{E}_{\mathbf{x} \sim \mathcal{D}_{\text{train}}} \left[\frac{1}{T} \sum_{t=1}^{T} \mathbf{x}_t \right] \tag{4}$$

$$\mathbf{s} = \sqrt{\frac{1}{N-1} \sum_{n \in \mathcal{D}_{\text{train}}} \sum_{t=1}^{T_n} (\mathbf{x}_t^{(n)} - \bar{\mathbf{x}})^2} \tag{5}$$

Data leakage experiment

We used a k nearest neighbours (KNN) classifier to investigate the inflationary effect of data leakage. Our KNN classifier assigned new data points by looking at the $k = 5$ most similar data points it had seen during training. It then chose the label the majority of the k most similar data points had. In other words, KNN classified purely on similarity, and to assess this similarity, KNN calculated a Euclidean distance $d(\mathbf{x}_i, \mathbf{x}_j)$ for two vector representations of audio:

$$d(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{\sum_{k=1}^{n} (x_{i_k} - x_{j_k})^2}, \tag{6}$$

with $n = 60$ in our case.

Our experiment comprised two conditions. In the non-leakage condition, we split recordings into two halves in the middle m with $m = \lfloor T/2 \rfloor$, creating two vectors with averaged MFCCs.

$$\mathbf{x}_L = \frac{1}{m} \sum_{t=1}^{m} \mathbf{x}_t \tag{7}$$

$$\mathbf{x}_R = \frac{1}{T-m} \sum_{t=m+1}^{T} \mathbf{x}_t \tag{8}$$

Once recordings were split, we then selected an independent train set and fed the left-hand side of the recording to the model. Then, we evaluated on the right-hand side of the recording of an independent test set. In other words, in this condition, our model never saw the two halves of any one person's recording but only one half of each person.

In the leakage condition, we still trained on the left-hand side of a recording and tested on the right-hand. However, we drew samples from the entire data set, not just one of the sets, to generate these halves. The number of samples that the model was evaluated on in this condition was the same as in the non-leakage condition.

This design also explains our choice of KNN for our leakage experiment: KNN isolates the causal mechanism. Any improvement under the leakage condition can be attributed directly to matching a speaker during evaluation who is similar to one seen during training. In other words, speakers are recognized [38].

For a variety of reasons, our experiment was deliberately conservative. First, we only used MFCCs as predictors, which only measured frequencies, rather than a multi predictor approach that might give a model more surface to recognize a speaker. Second, we mitigated idiosyncratic frequencies speakers might have exhibited, because we used the rainbow task as a source of MFCCs, in which everyone spoke the exact same words. Third, we only split a single recording into two halves and averaged MFCCs across these two halves. In contrast, we could have split recordings into multiple segments in which many shorter, adjacent segments are more likely to be similar to one another. This would have given our model more material to recognize individual speakers. The reason short, adjacent segments are more similar to one another, rather than two long halves, is auto-correlation: speech is a continuous wave; physically it is vibration of air molecules. MFCCs that are only a short distance δ apart are more correlated than MFCCs that are more timesteps Δ apart, simply because speech changes gradually.

$$\mathbb{E}\left[d(\mathbf{x}_t, \mathbf{x}_{t+\delta})^2\right] \leq \mathbb{E}\left[d(\mathbf{x}_t, \mathbf{x}_{t+\Delta})^2\right] \quad \text{for } \delta < \Delta \tag{9}$$

Correlation analysis

Based on literature, we selected several predictors a priori that were supposed to show strong correlations with our outcomes [42], [43]. Since much of the existing literature is based on small samples, we interpreted these prior findings with caution. The predictors we chose, in the end, were the normalised standard deviation of the first four MFCCs; the mean F0, F1, and F2 frequencies; the mean cepstral peak prominence; the local jitter; the mean harmonics-to-noise ratio; and the speaking rate. We then calculated Spearman's correlation coefficients for these variables and our binary outcomes. We used Spearman's rather than Pearson's correlation coefficient, as our outcomes were ordinal variables.

Static classifiers

We created three different classifiers as they seemed to have done well in previous research: a logistic regression, a support vector machine, and a random forest classifier [10]. We did not conduct any hyperparameter searches. Our logistic regression used a pure L2 penalty with a liblinear solver. Our random forest classifier had 100 trees and required at least 50 samples to split an internal node, and at least 20 samples for a node to be a leaf. The support vector machine used a radial basis function kernel, a full L2 penalty, and a scaled gamma coefficient.

We used speaker-independent train-test splits across our seeds. The ratio was always 80:20. We weighed data by inverse class frequency to account for imbalances. We implemented the static classifiers in scikit-learn.

Time-series models

We used a unidirectional, long short-term memory (LSTM) classifier with a single recurrent layer to predict mental health outcomes from time-series speech data [56]. We chose an LSTM, as there is a belief that it is particularly well-suited to the prediction of mental disorders [9]. Unlike standard neural networks that cannot natively process time-series data, LSTMs can retain a memory of the time-series, selectively including or discarding prior information [56]. Consequently, the hope behind the use of an LSTM is that the model yields improved classification at the final timestep T of a recording, because it has learned to integrate temporally relevant patterns from preceding timesteps $t < T$. An alternative to our LSTM would have been a transformer architecture. We chose the LSTM, because LSTMs seem competitive with transformers [57] and may even outperform them for time-series data [58].

We describe below how our LSTMs worked in practice. The LSTMs retained and discarded information by maintaining a cell state c_t, which was updated with the MFCC feature vector at each time step $\mathbf{x}_t \in \mathbb{R}^{60}$.

What information flowed in and out of the cell states, specifically, was controlled via the previous cell state c_{t-1}, the forget gate f_t, the input gate i_t, and the candidate cell-state vector g_t.

$$c_t = f_t \odot c_{t-1} + i_t \odot g_t, \tag{10}$$

The input, forget, and candidate cell-state vectors were computed by applying nonlinear activation functions to affine transformations of the MFCC input $\mathbf{x_t}$ and the previous hidden state $\mathbf{h}_{t-1}$.

$$i_t = \sigma(W_{ii}x_t + b_{ii} + W_{hi}h_{t-1} + b_{hi}) \tag{11}$$

$$f_t = \sigma(W_{if}x_t + b_{if} + W_{hf}h_{t-1} + b_{hf}) \tag{12}$$

$$g_t = \tanh(W_{ig}x_t + b_{ig} + W_{hg}h_{t-1} + b_{hg}) \tag{13}$$

The hidden vector of $h_t \in \mathbb{R}^{32}$ was then computed via another vector o_t:

$$o_t = \sigma(W_{io}x_t + b_{io} + W_{ho}h_{t-1} + b_{ho}) \tag{14}$$

$$h_t = o_t \odot \tanh(c_t) \tag{15}$$

We then passed the final hidden vector h_T through a linear layer to obtain a classification of the recording. We used a cross-entropy loss function, and we used Adam with $\alpha = 10^{-3}$ for stochastic optimisation [59].

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)] \tag{16}$$

We trained for ten epochs. As before, we used speaker-independent train-test splits across seeds, with a ratio of 70:30. We weighted data in the same way as with our static classifiers. We did not tune hyperparameters, and we implemented the LSTMs in PyTorch.

References

[1] GBD 2019 Mental Disorders Collaborators, "Global, regional, and national burden of 12 mental disorders in 204 countries and territories, 1990-2019: A systematic analysis for the Global Burden of Disease Study 2019," *The Lancet. Psychiatry*, vol. 9, no. 2, pp. 137–150, Feb. 2022, doi: 10.1016/S2215-0366(21)00395-3.

[2] D. M. Clark, "Realizing the Mass Public Benefit of Evidence-Based Psychological Therapies: The IAPT Program," *Annual Review of Clinical Psychology*, vol. 14, pp. 159–183, May 2018, doi: 10.1146/annurev-clinpsy-050817-084833.

[3] M. P. Acuña-Rodríguez, O. Fiorillo-Moreno, K. F. Montoya-Quintero, and F. Tejan Mansaray, "Mental Health Workforce Inequities Across Income Levels: Aligning Global Health Indicators, Policy Readiness, and Disease Burden," *Psychology Research and Behavior Management*, vol. 18, pp. 1449–1454, 2025, doi: 10.2147/PRBM.S532912.

[4] T. Gagné, C. Henderson, and A. McMunn, "Is the self-reporting of mental health problems sensitive to public stigma towards mental illness? A comparison of time trends across English regions (2009–19)," *Social Psychiatry and Psychiatric Epidemiology*, vol. 58, no. 4, pp. 671–680, Apr. 2023, doi: 10.1007/s00127-022-02388-7.

[5] S. Newman and V. G. Mather, "ANALYSIS OF SPOKEN LANGUAGE OF PATIENTS WITH AFFECTIVE DISORDERS," *American Journal of Psychiatry*, vol. 94, no. 4, pp. 913–942, Jan. 1938, doi: 10.1176/ajp.94.4.913.

[6] M. Wiseman *et al.*, "Objective speech measures capture depressive symptoms and associated cognitive difficulties," *Translational Psychiatry*, vol. 15, no. 1, p. 525, Nov. 2025, doi: 10.1038/s41398-025-03728-2.

[7] Y. Li, S. Kumbale, Y. Chen, T. Surana, E. S. Chng, and C. Guan, "Automated Depression Detection From Text and Audio: A Systematic Review," *IEEE Journal of Biomedical and Health Informatics*, vol. 29, no. 10, pp. 7498–7513, Oct. 2025, doi: 10.1109/JBHI.2025.3570900.

[8] S. Kapoor and A. Narayanan, "Leakage and the reproducibility crisis in machine-learning-based science," *Patterns*, vol. 4, no. 9, p. 100804, Sep. 2023, doi: 10.1016/j.patter.2023.100804.

[9] L. Liu, L. Liu, H. A. Wafa, F. Tydeman, W. Xie, and Y. Wang, "Diagnostic accuracy of deep learning using speech samples in depression: A systematic review and meta-analysis," *Journal of the American Medical Informatics Association*, vol. 31, no. 10, pp. 2394–2404, Oct. 2024, doi: 10.1093/jamia/ocae189.

[10] B. S, D. S. Nayak, R. C. Dmello, A. Nayak, and S. S. Bangera, "Machine Learning Applied to Speech Emotion Analysis for Depression Recognition," in *2023 International Conference for Advancement in Technology (ICONAT)*, Goa, India: IEEE, Jan. 2023, pp. 1–5. doi: 10.1109/ICONAT57137.2023.10080060.

[11] K. Chlasta, K. Wołk, and I. Krejtz, "Automated speech-based screening of depression using deep convolutional neural networks," *Procedia Computer Science*, vol. 164, pp. 618–628, 2019, doi: 10.1016/j.procs.2019.12.228.

[12] Y. Cui, Z. Li, L. Liu, J. Zhang, and J. Liu, "Privacy-preserving Speech-based Depression Diagnosis via Federated Learning," in *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, Glasgow, Scotland, United Kingdom: IEEE, Jul. 2022, pp. 1371–1374. doi: 10.1109/EMBC48229.2022.9871861.

[13] A. K. Das and R. Naskar, "A deep learning model for depression detection based on MFCC and CNN generated spectrogram features," *Biomedical Signal Processing and Control*, vol. 90, p. 105898, Apr. 2024, doi: 10.1016/j.bspc.2023.105898.

[14] M. Du *et al.*, "Depression recognition using a proposed speech chain model fusing speech production and perception features," *Journal of Affective Disorders*, vol. 323, pp. 299–308, Feb. 2023, doi: 10.1016/j.jad.2022.11.060.

[15] K. Feng and T. Chaspari, "A Knowledge-Driven Vowel-Based Approach of Depression Classification from Speech Using Data Augmentation," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Rhodes Island, Greece: IEEE, Jun. 2023, pp. 1–5. doi: 10.1109/ICASSP49357.2023.10096105.

[16] S. Gupta, G. Agarwal, S. Agarwal, and D. Pandey, "Depression detection using cascaded attention based deep learning framework using speech data," *Multimedia Tools and Applications*, vol. 83, no. 25, pp. 66135–66173, Jan. 2024, doi: 10.1007/s11042-023-18076-w.

[17] P. Homsiang, T. Treebupachatsakul, K. Kiatrungrit, and S. Poomrittigul, "Classification of Depression Audio Data by Deep Learning," in *2022 14th Biomedical Engineering International Conference (BMEiCON)*, Songkhla, Thailand: IEEE, Nov. 2022, pp. 1–4. doi: 10.1109/BMEiCON56653.2022.10012102.

[18] M. Ishimaru, Y. Okada, R. Uchiyama, R. Horiguchi, and I. Toyoshima, "Classification of Depression and Its Severity Based on Multiple Audio Features Using a Graphical Convolutional Neural Network," *International Journal of Environmental Research and Public Health*, vol. 20, no. 2, p. 1588, Jan. 2023, doi: 10.3390/ijerph20021588.

[19] A. Z. Jenei and G. Kiss, "Possibilities of Recognizing Depression with Convolutional Networks Applied in Correlation Structure," in *2020 43rd International Conference on Telecommunications and Signal Processing (TSP)*, Milan, Italy: IEEE, Jul. 2020, pp. 101–104. doi: 10.1109/TSP49548.2020.9163547.

[20] A. Z. Jenei and G. Kiss, "Severity Estimation of Depression Using Convolutional Neural Network," *Periodica Polytechnica Electrical Engineering and Computer Science*, vol. 65, no. 3, pp. 227–234, Jun. 2021, doi: 10.3311/PPee.15958.

[21] Vandana, N. Marriwala, and D. Chaudhary, "A hybrid model for depression detection using deep learning," *Measurement: Sensors*, vol. 25, p. 100587, Feb. 2023, doi: 10.1016/j.measen.2022.100587.

[22] A. Othmani, D. Kadoch, K. Bentounes, E. Rejaibi, R. Alfred, and A. Hadid, "Towards Robust Deep Neural Networks for Affect and Depression Recognition from Speech," in *Pattern Recognition. ICPR International Workshops and Challenges*, vol. 12662, A. Del Bimbo, R. Cucchiara, S. Sclaroff, G. M. Farinella, T. Mei, M. Bertini, H. J. Escalante, and R. Vezzani, Eds., Cham: Springer International Publishing, 2021, pp. 5–19. doi: 10.1007/978-3-030-68790-8_1.

[23] V. Ravi, J. Wang, J. Flint, and A. Alwan, "A Step Towards Preserving Speakers' Identity While Detecting Depression Via Speaker Disentanglement," in *Interspeech 2022*, ISCA, Sep. 2022, pp. 3338–3342. doi: 10.21437/Interspeech.2022-10798.

[24] V. Ravi, J. Wang, J. Flint, and A. Alwan, "Enhancing accuracy and privacy in speech-based depression detection through speaker disentanglement," *Computer Speech & Language*, vol. 86, p. 101605, Jun. 2024, doi: 10.1016/j.csl.2023.101605.

[25] E. Rejaibi, A. Komaty, F. Meriaudeau, S. Agrebi, and A. Othmani, "MFCC-based Recurrent Neural Network for automatic clinical depression recognition and assessment from speech," *Biomedical Signal Processing and Control*, vol. 71, p. 103107, Jan. 2022, doi: 10.1016/j.bspc.2021.103107.

[26] A. Saidi, S. B. Othman, and S. B. Saoud, "Hybrid CNN-SVM classifier for efficient depression detection system," in *2020 4th International Conference on Advanced Systems and Emergent Technologies (IC_ASET)*, Hammamet, Tunisia: IEEE, Dec. 2020, pp. 229–234. doi: 10.1109/IC_ASET49463.2020.9318302.

[27] S. Sardari, B. Nakisa, M. N. Rastgoo, and P. Eklund, "Audio based depression detection using Convolutional Autoencoder," *Expert Systems with Applications*, vol. 189, p. 116076, Mar. 2022, doi: 10.1016/j.eswa.2021.116076.

[28] S. Suparatpinyo and N. Soonthornphisaj, "Smart voice recognition based on deep learning for depression diagnosis," *Artificial Life and Robotics*, vol. 28, no. 2, pp. 332–342, May 2023, doi: 10.1007/s10015-023-00852-4.

[29] H. Tian, Z. Zhu, and X. Jing, "Deep learning for Depression Recognition from Speech," *Mobile Networks and Applications*, vol. 29, no. 4, pp. 1212–1227, Aug. 2024, doi: 10.1007/s11036-022-02086-3.

[30] Q. Wang and N. Liu, "Speech Detection of Depression Based on Multi-mlp," in *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Las Vegas, NV, USA: IEEE, Dec. 2022, pp. 3896–3898. doi: 10.1109/BIBM55620.2022.9995447.

[31] Y. Wang, X. Lu, and D. Shi, "MFCC-based deep convolutional neural network for audio depression recognition," in *2022 International Conference on Asian Language Processing (IALP)*, Singapore, Singapore: IEEE, Oct. 2022, pp. 162–166. doi: 10.1109/IALP57159.2022.9961267.

[32] W. Yang *et al.*, "Attention guided learnable time-domain filterbanks for speech depression detection," *Neural Networks*, vol. 165, pp. 135–149, Aug. 2023, doi: 10.1016/j.neunet.2023.05.041.

[33] F. Yin, J. Du, X. Xu, and L. Zhao, "Depression Detection in Speech Using Transformer and Parallel Convolutional Neural Networks," *Electronics*, vol. 12, no. 2, p. 328, Jan. 2023, doi: 10.3390/electronics12020328.

[34] Z. Zhou, Y. Guo, S. Hao, and R. Hong, "Hierarchical Multifeature Fusion via Audio-Response-Level Modeling for Depression Detection," *IEEE Transactions on Computational Social Systems*, vol. 10, no. 5, pp. 2797–2805, Oct. 2023, doi: 10.1109/TCSS.2022.3202294.

[35] F. Ringeval *et al.*, "AVEC 2018 Workshop and Challenge: Bipolar Disorder and Cross-Cultural Affect Recognition," in *Proceedings of the 2018 on Audio/Visual Emotion Challenge and Workshop*, Seoul Republic of Korea: ACM, Oct. 2018, pp. 3–13. doi: 10.1145/3266302.3266316.

[36] L. Manea, S. Gilbody, and D. McMillan, "Optimal cut-off score for diagnosing depression with the Patient Health Questionnaire (PHQ-9): A meta-analysis," *Canadian Medical Association Journal*, vol. 184, no. 3, pp. E191–E196, Feb. 2012, doi: 10.1503/cmaj.110829.

[37] M. Rosenblatt, L. Tejavibulya, R. Jiang, S. Noble, and D. Scheinost, "Data leakage inflates prediction performance in connectome-based machine learning models," *Nature Communications*, vol. 15, no. 1, p. 1829, Feb. 2024, doi: 10.1038/s41467-024-46150-w.

[38] H.-C. Yeh, L. Sun, A. Mahapatra, S. S. Chandra, E. M. Provost, and B. Sisman, "Who is Speaking or Who is Depressed? A Controlled Study of Speaker Leakage in Speech-Based Depression Detection." arXiv, Apr. 2026. doi: 10.48550/arXiv.2604.14354.

[39] D. Picard, "Torch.manual_seed(3407) is all you need: On the influence of random seeds in deep learning architectures for computer vision." arXiv, May 2023. doi: 10.48550/arXiv.2109.08203.

[40] L. Schader, W. Song, R. Kempker, and D. Benkeser, "Don't Let Your Analysis Go to Seed: On the Impact of Random Seed on Machine Learning-based Causal Inference," *Epidemiology*, vol. 35, no. 6, pp. 764–778, Nov. 2024, doi: 10.1097/EDE.0000000000001782.

[41] Y. Bensoussan *et al.*, "Bridge2AI-Voice: An ethically-sourced, diverse voice dataset linked to health information." PhysioNet. doi: 10.13026/3XT6-RF05.

[42] Y. Yang, C. Fairbairn, and J. F. Cohn, "Detecting Depression Severity from Vocal Prosody," *IEEE Transactions on Affective Computing*, vol. 4, no. 2, pp. 142–150, Apr. 2013, doi: 10.1109/T-AFFC.2012.38.

[43] M.-S. Seifpanahi, T. Ghaemi, A. Ghaleiha, D. Sobhani-Rad, and M.-K. Zarabian, "The Association between Depression Severity, Prosody, and Voice Acoustic Features in Women with Depression," *The Scientific World Journal*, vol. 2023, pp. 1–8, Dec. 2023, doi: 10.1155/2023/9928446.

[44] F. Eyben *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, Apr. 2016, doi: 10.1109/TAFFC.2015.2457417.

[45] F. D. Schönbrodt and M. Perugini, "At what sample size do correlations stabilize?" *Journal of Research in Personality*, vol. 47, no. 5, pp. 609–612, Oct. 2013, doi: 10.1016/j.jrp.2013.05.009.

[46] O. Colliot, E. Thibeau-Sutre, and N. Burgos, "Reproducibility in machine learning for medical imaging." arXiv, 2022. doi: 10.48550/ARXIV.2209.05097.

[47] A. Ramaswamy *et al.*, "ChatGPT Health performance in a structured test of triage recommendations," *Nature Medicine*, vol. 32, no. 5, pp. 1671–1675, May 2026, doi: 10.1038/s41591-026-04297-7.

[48] G. Varoquaux and V. Cheplygina, "Machine learning for medical imaging: Methodological failures and recommendations for the future," *npj Digital Medicine*, vol. 5, no. 1, p. 48, Apr. 2022, doi: 10.1038/s41746-022-00592-y.

[49] W. J. Youden, "Index for rating diagnostic tests," *Cancer*, vol. 3, no. 1, pp. 32–35, 1950, doi: 10.1002/1097-0142(1950)3:1<32::AID-CNCR2820030106>3.0.CO;2-3.

[50] C. Colas, O. Sigaud, and P.-Y. Oudeyer, "How Many Random Seeds? Statistical Power Analysis in Deep Reinforcement Learning Experiments." arXiv, Jul. 2018. doi: 10.48550/arXiv.1806.08295.

[51] A. Johnson *et al.*, "Bridge2AI-Voice: An ethically-sourced, diverse voice dataset linked to health information." healthdatanexus.ai. doi: 10.57764/QB6H-EM84.

[52] K. Kroenke, R. L. Spitzer, and J. B. W. Williams, "The PHQ-9: Validity of a brief depression severity measure," *Journal of General Internal Medicine*, vol. 16, no. 9, pp. 606–613, Sep. 2001, doi: 10.1046/j.1525-1497.2001.016009606.x.

[53] A. Rameau *et al.*, "Developing Multi-Disorder Voice Protocols: A team science approach involving clinical expertise, bioethics, standards, and DEI." in *Interspeech 2024*, ISCA, Sep. 2024, pp. 1445–1449. doi: 10.21437/Interspeech.2024-1926.

[54] S. Alghowinem, R. Goecke, M. Wagner, J. Epps, M. Breakspear, and G. Parker, "Detecting depression: A comparison between spontaneous and read speech," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, Vancouver, BC, Canada: IEEE, May 2013, pp. 7547–7551. doi: 10.1109/ICASSP.2013.6639130.

[55] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 4, pp. 357–366, Aug. 1980, doi: 10.1109/TASSP.1980.1163420.

[56] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997, doi: 10.1162/neco.1997.9.8.1735.

[57] K. Hasan, J. Saquer, and M. Ghosh, "Advancing Mental Disorder Detection: A Comparative Evaluation of Transformer and LSTM Architectures on Social Media." arXiv, Jul. 2025. doi: 10.48550/arXiv.2507.19511.

[58] A. Zeng, M. Chen, L. Zhang, and Q. Xu, "Are Transformers Effective for Time Series Forecasting?" arXiv, Aug. 2022. doi: 10.48550/arXiv.2205.13504.

[59] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization." arXiv, Jan. 2017. doi: 10.48550/arXiv.1412.6980.