Exploration of Perceptual Speech Features for Clinical Decision-Support in Mental Health Care

Vassilis Lyberatos[1,2], Edmund G. Dervakos[2], Eleni Adamidi[2], Athanasios Voulodimos[1], Giorgos Stamou[1]

[1]National Technical University of Athens, Athens, Greece
[2]PsychNow
vaslyb@ails.ece.ntua.gr
eddie@psychnow.com

Abstract

Speech and language technologies offer valuable opportunities for supporting mental health assessment through objective and interpretable cues. We present a systematic feature-based analysis framework leveraging perceptually grounded acoustic and linguistic characteristics, including prosody, vocal quality, semantic coherence, syntactic structure, and sarcasm. Using statistical analysis and interpretable machine learning (XGBoost with SHAP and LIME), we examine associations between speech features and validated symptom measures of depression, anxiety, and ADHD. Evaluated on both controlled benchmark datasets (StressID, DAIC-WOZ, Androids, EATD) and a real-world clinical dataset, the framework reveals stable and consistent relationships between symptom severity and vocal irregularities (e.g., shimmer, jitter), lexical-syntactic patterns, and affective tone. An ablation study conducted across all datasets further identifies the most informative feature groups. This work explores a transparent and clinically interpretable approach to speech-based mental health analysis.

1 Introduction

Mental health disorders represent a major global health burden, affecting nearly 970 million people worldwide in 2019 and accounting for approximately 16% of global years lived with disability (YLDs), which makes them one of the leading causes of disability worldwide (Vos et al., 2020). Traditional screening and intake approaches rely heavily on subjective evaluations, including clinical interviews and self-reported symptoms, which are time-consuming and vulnerable to clinician bias, recall bias, and stigma-driven underreporting. In this context, there is a pressing need not for diagnostic tools in the narrow sense, but for clinically supportive technologies that provide objective and interpretable cues without enforcing rigid categorical decisions (Berisha and Liss, 2024). Particularly in the early stages of assessment, tools that function as perceptual and behavioral aids can help clinicians detect patterns, guide conversation, and observe phenotypes while avoiding premature labeling and stigma reinforcement (Kotov et al., 2017).

Speech and language provide a rich, non-invasive source of information for understanding mental health. As natural modes of human expression, they encode cognitive, emotional, and neurological states through acoustic properties (e.g., pitch, prosody, intensity) and linguistic structure (e.g., lexical richness, syntactic complexity) (Cummins et al., 2015, 2018). A growing body of research shows that psychiatric disorders are reflected in distinctive speech and language patterns, including alterations in articulation, semantic coherence, prosody, and syntax, particularly in conditions such as depression, anxiety, schizophrenia, and cognitive impairment (Alhanai et al., 2018; Arevian et al., 2020).

Recent advances in speech and language technologies have enabled automated analysis for a range of mental health applications. Speech collected in naturalistic settings, including mobile environments, has been used to identify depression, anxiety, insomnia, and fatigue with encouraging accuracy and robustness across populations (Riad et al., 2024). However, many state-of-the-art models function as "black boxes," which limits interpretability and undermines clinician trust. There is growing demand, motivated by both ethical and regulatory considerations, for explainable and interpretable AI systems in healthcare, particularly in high-stakes domains such as mental health assessment (Ng et al., 2024; Holzinger et al., 2019; Doshi-Velez and Kim, 2017).

Perceptually motivated acoustic and linguistic features, such as prosodic, spectral, and lexical-syntactic markers, are inherently interpretable and grounded in well-established clinical phenomena

(Tasnim et al., 2022; Voleti et al., 2019; Jiao et al., 2017; Tu et al., 2017; Premananth et al., 2025; Neumann et al., 2025). Models that rely on these features are typically more transparent and accessible to clinicians. Such features can be obtained through standard signal processing, extracted using auxiliary deep neural networks (Jiao et al., 2017; Tu et al., 2017), or learned directly in end-to-end architectures (Leschly et al., 2025; Korzekwa et al., 2019; Xu et al., 2023). When combined with interpretable classifiers or post-hoc explanation methods such as attribution analyses, perceptual feature pipelines offer clear insight into which speech characteristics drive model predictions. This approach improves accountability and supports clinical decision-making by providing human-understandable evidence that aligns with expert knowledge. Consequently, perceptual feature-based explainable AI frameworks are a promising strategy for bridging complex machine learning models and practical mental health assessment (Ntalampiras, 2025; Menne et al., 2024).

In this work, we explore the development of transparent and clinically meaningful speech and language analysis methods for mental health assessment. We present a systematic, interpretable framework that combines perceptually grounded acoustic and linguistic features with feature-based modeling and explainable AI techniques (Guidotti et al., 2018). To evaluate the robustness and generalizability of the proposed approach, we conduct experiments on five speech datasets spanning controlled laboratory stress elicitation, semi-structured clinical interviews, multilingual public depression corpora, and real-world digital mental health assessments. Across datasets, the framework is applied consistently to binary classification tasks involving stress, depression, anxiety, and attention-related difficulties. In addition, we perform statistical analyses and ablation studies on feature aggregation and representation strategies to assess the stability and interpretability of the extracted speech markers. By examining speech characteristics across diverse recording conditions, languages, and clinical contexts, this work explores a clinically grounded approach that links perceptual feature design with explainable modeling, contributing to more transparent and interpretable speech-based technologies for mental health.

2 Methodology

Our guiding principle in designing the experimental methodology was to prioritize the extraction of clinically interpretable features by leveraging reliable tools. In line with prior work on clinical applications of voice analysis (Jiao et al., 2017; Riad et al., 2024; Donnelly et al., 2024), we adopted feature extraction strategies that draw upon multiple disciplines, including natural language processing, signal processing, and auxiliary deep neural models. As a next step, we applied traditional statistical and machine learning analyses, further enhanced with post-hoc explainable AI (XAI) techniques (Guidotti et al., 2018), in order to improve the interpretability of our results and analysis.

2.1 Feature Extraction

Clinical psychopathology can be detected from speech based not only on what is said, but also on how it is said (Aloshban et al., 2022). In our approach, we incorporate both modalities, audio and language, into the analysis. Accordingly, our feature extraction methods can be divided into two categories: acoustic features and linguistic features. A list of all extracted features is provided in Table 5 in Appendix D.

2.1.1 Acoustic Features

For the audio analysis, we used Parselmouth (Jadoul et al., 2018), which enables Praat 's (Boersma and Weenink, 2021) validated acoustic analyses directly within python. Acoustic features were extracted from voiced segments and included pitch statistics, intensity, jitter, shimmer, harmonic-to-noise ratio (HNR), zero-crossing rate (ZCR), pauses, phonation and articulation rates, rhythm (pairwise variability index), and speech entropy. Prior to analysis, waveforms were converted to mono, resampled to 16 kHz, and normalized in amplitude to ensure comparability across recordings. Beyond low-level acoustics, we extracted higher-level paralinguistic and linguistic representations using pretrained neural models. Emotion-related features were obtained with a HuBERT-based speech emotion recognition model [1] (Yang et al., 2021), fine-tuned on the IEMOCAP corpus (Busso et al., 2008). The extracted acoustic descriptors are organized into three interpretable groups: prosodic/fluency

[1] https://huggingface.co/superb/hubert-base-superb-er

features (e.g., pitch and intensity statistics, pause measures, phonation and articulation rates, and rhythm/variability indices), voice quality features (e.g., jitter, shimmer, and harmonic-to-noise ratio), and psycholinguistic (auxiliary emotion and sarcasm estimates).

These representations have also been studied in prior clinical work. Prosodic reductions such as flatter pitch range/variability and increased pausing are commonly associated with monotone speech, depression, and blunted affect, while elevated variability may reflect agitation or manic states (Alpert et al., 2001; Low et al., 2020). Voice quality perturbations have been linked to psychopathology in prior work, with shimmer in particular reported as a correlate of depression severity in some settings (Ettore et al., 2022; Hönig et al., 2014). Finally, sarcasm and related pragmatic indicators have been associated with increased risk of anxiety, stress, and depression (Dionigi et al., 2023; Gross and Jazaieri, 2014; Pope et al., 1970).

2.1.2 Linguistic Features

Linguistic features were extracted from transcripts using spaCy (Honnibal et al., 2020) and Stanza (Qi et al., 2020), providing tokenization, POS tagging, lemmatization, dependency parsing, and constituency trees. From these annotations, we derived lexical indices (type-token ratio, MATTR, Brunet's index, Honore's statistic, lemma diversity, and morphological richness), syntactic measures (mean sentence length, clause ratio, syntactic and constituency depth, and passive voice ratio), and graph-based discourse metrics (connectivity, loops, density, diameter, and path statistics). Psycholinguistic information was obtained using VADER sentiment analysis (Hutto and Gilbert, 2014), capturing positive, neutral, and negative valence. This set of descriptors captures lexical diversity, morphosyntactic complexity, discourse organization, and affective tone. To further capture semantic information, pretrained neural models were integrated, with sentence-level embeddings obtained using Sentence-BERT [2] (Reimers and Gurevych, 2019), enabling estimation of discourse coherence, cohesion, and repetition patterns. The resulting linguistic descriptors are organized into four interpretable groups: lexical features (e.g., word/sentence counts, lexical diversity indices such as TTR/MATTR/Brunet/Honoré, content–function and pronoun ratios, and morphology/POS diversity), syntactic features (e.g., sentence/clause lengths, dependency and constituency depth, passive voice usage, and graph-based measures of discourse structure and repetition), semantic features (e.g., first-/second-order coherence and repetition/cohesion measures), and psycholinguistic features capturing affective tone and pragmatics, including sentiment and sarcasm.

These groupings are motivated by prior clinical findings: reduced lexical richness and shifts in tense/pronoun usage have been linked to schizophrenia, dementia, and depression, and pronoun patterns may reflect self-focus or social withdrawal (Compton et al., 2023; Pennebaker et al., 2003). Reduced syntactic complexity is commonly associated with cognitive impairment and depression, while increased repetition and disfluency-like patterns have been reported in ADHD (Sung et al., 2020; Engelhardt et al., 2011). At the semantic level, reduced coherence has been associated with disorganized thought and psychosis-related phenomena and has also been observed in ADHD and manic states (Corcoran et al., 2018; Engelhardt et al., 2011). Finally, affective language indicators (e.g., elevated negative sentiment and reduced positive emotion) are associated with mood disorders, while pragmatic cues such as sarcasm can correlate with elevated risk and may track agitation/psychosis-related expression in some settings (Sonnenschein et al., 2018; Dionigi et al., 2023).

2.1.3 Sarcasm

Sarcasm is a subtle communicative cue that conveys implicit emotional and attitudinal meaning beyond literal word content. Since sarcasm has been correlated with depressive symptoms (Dionigi et al., 2023), it provides a relevant paralinguistic marker for our study. To capture this dimension, we trained a multimodal sarcasm detection model on the MUStARD dataset (Castro et al., 2019), achieving an accuracy of approximately 70%. The model integrates linguistic and acoustic information by combining BERT[3] (Devlin et al., 2019) for text and Wav2Vec2[4] (Baevski et al., 2020) for audio, with frozen encoders whose pooled embeddings are projected into a shared space, concate-

[2] https://huggingface.co/sentence-transformers/paraphrase-MiniLM-L6-v2

[3] https://huggingface.co/google-bert/bert-base-uncased

[4] https://huggingface.co/facebook/wav2vec2-base-960h

nated, and passed through a feedforward classifier. Once trained, this auxiliary model was applied to our datasets to infer sarcasm probabilities, which were then included as additional features alongside acoustic, linguistic, and emotional representations to enrich our analysis. This procedure yielded a total of 82 scalar, interpretable features.

2.2 Analytical Framework

The analytical framework was designed to explore statistical and structural relationships between psychopathology and speech features in an exploratory manner. We combined inferential statistics with interpretable machine learning and explainability techniques, an integration that, to our knowledge, remains relatively uncommon in this area, to examine associations across datasets.

Group-level comparisons were first conducted using independent samples t-tests to assess significant differences in perceptual and paralinguistic features between participant subgroups defined by clinical thresholds on PHQ-9, GAD-7, and ASRS scores. To control for multiple comparisons, resulting p-values were adjusted using the false discovery rate (FDR) correction according to the Benjamini-Hochberg procedure. These tests provided an initial understanding of feature distributions and effect directions underlying symptom-related variation.

To capture higher-order and nonlinear dependencies, we employed XGBoost (Chen and Guestrin, 2016) classifiers as analytical models linking voice features to mental health categories. XGBoost was selected due to its strong performance on tabular data and its compatibility with transparent, feature-level interpretability. The internal structure of these models was examined through feature importance rankings and post-hoc explainability analyses using SHAP (Lundberg and Lee, 2017) and LIME (Ribeiro et al., 2016). LIME explanations were aggregated across all instances to derive cumulative patterns of feature influence, enabling a global interpretation of localized explanations. Complementarily, SHAP summary plots were employed to visualize the overall distribution, magnitude, and direction of feature effects. Together, these analyses indicated correlations and interaction patterns between prosodic, spectral, and temporal features and psychopathological dimensions.

In addition, we conducted an ablation study to identify the most informative groups of features, using the feature groupings defined in the Section 2.1. By systematically training models with individual feature groups isolated, we assessed the relative contribution of acoustic, prosodic, spectral, and linguistic feature sets to the observed associations with psychopathology. This analysis provided further insight into which categories of features drive the explanatory power of the framework.

Overall, this framework follows an explanatory modeling approach in which both statistical tests and interpretable machine learning are used to identify and validate relationships between acoustic patterns and mental health indicators, providing analytical insight rather than purely predictive outcomes.

3 Experiments

We conducted experiments on five speech datasets spanning controlled laboratory conditions, semi-structured clinical interviews, multilingual public corpora, and real-world digital mental health assessments. Across all datasets, we applied the proposed feature extraction pipeline to extract acoustic and linguistic descriptors from speech and examined their relationship to clinically meaningful labels and screening instruments. Depending on the dataset, the experimental task was formulated as binary classification, including stress recognition and the detection of depression, anxiety, and attention-related difficulties. Statistical analysis and explainable machine learning methods were used to support robust evaluation and interpretable feature attribution.

3.1 Datasets

As a primary benchmark, we used the StressID dataset (Chaptoukaev et al., 2023), a multimodal corpus designed for stress identification. StressID includes synchronized audio, facial video, and physiological recordings (ECG, EDA, respiration) from 65 participants performing 11 stress-inducing and neutral tasks. The dataset contains over 39 hours of annotated data, with self-reported measures of stress, relaxation, arousal, and valence. In this study, we used only the audio modality and formulated a binary classification task distinguishing stressed from non-stressed speech.

To assess depression detection in structured clinical interviews, we used the DAIC-WOZ corpus (Gratch et al., 2014), a subset of the Distress Analysis Interview Corpus. It consists of semi-structured interviews conducted by a virtual agent using a Wizard-of-Oz framework, and includes synchronized audio, video, transcripts, and validated

PHQ-8 depression scores. We extracted participant speech only and performed binary depression classification based on standard PHQ-8 thresholds.

We further evaluated our approach on the ANDROIDS corpus (Tao et al., 2023), a clinically validated Italian-language dataset comprising speech recordings from 118 participants, including individuals diagnosed with depression and healthy controls. The dataset contains both read and spontaneous speech collected in real clinical environments, with expert diagnostic labels. In our experiments, we used the audio recordings to perform binary depression classification.

To examine cross-linguistic generalization, we employed the EATD corpus (Shen et al., 2022), the first publicly available Chinese-language dataset for depression detection. EATD includes audio recordings and transcripts from 162 participants who completed standardized self-report assessments using the Self-Rating Depression Scale (SDS). Participants responded to emotionally eliciting prompts, and we used the speech recordings to define a binary depression classification task based on SDS score thresholds.

Finally, we used a proprietary real-world dataset collected through a digital mental health platform, comprising approximately 200 participants who completed a 30-minute self-report questionnaire and provided spoken responses, as part of psychiatric intake. The dataset includes audio recordings and standardized clinical screening scores for depression (PHQ-9), anxiety (GAD-7), and attention-deficit/hyperactivity disorder (ASRS). We analyzed speech features in relation to these clinical scores to evaluate the generalizability of our approach in real-world conditions. The dataset is balanced across diagnostic categories and participant sex. All participants provided informed consent, and all data were de-identified prior to analysis.

3.2 Results

Results are reported across five datasets (STRESSID, DAIC-WOZ, ANDROIDS, EATD, and REAL), spanning multiple clinical conditions, languages, and recording contexts. All experiments were formulated as binary classification tasks using dataset-specific labels or clinically validated questionnaire cutoffs.

The analyses focus on predictive performance using interpretable models, feature-group ablation to assess the contribution of predefined acoustic and linguistic representations, group-level statistical differences, and feature-level interpretability using SHAP, LIME, and partial dependence analysis. Additional results are provided in the Appendix F.

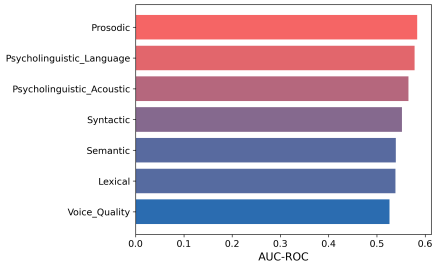


Figure 1: Average AUC-ROC across datasets for XGBoost models trained with only one feature group at a time.

3.2.1 Feature Group Ablation

An ablation study evaluated every predefined feature groups within each dataset. For each combination, a summary AUC-ROC was computed from cross-validation. Figure 1 reports the mean single-group AUC-ROC across datasets for each feature group in isolation. Prosodic features show the highest standalone average performance, followed by psycholinguistic language and acoustic groups. Syntactic, semantic, and lexical groups are intermediate, while voice-quality features are weakest alone. This cross-dataset summary suggests that no single group is sufficient, motivating the use of complementary feature combinations.

3.2.2 StressID

In the STRESSID dataset, we used the labels corresponding to the stress and non-stress conditions, with the dataset being balanced. In the original study (Chaptoukaev et al., 2023), which employed a large pretrained Wav2Vec model followed by a logistic regression classifier, an average accuracy of 0.66 (variance = 0.03) and an F1-score of 0.70 (variance = 0.02) were reported. In our approach, using perceptual features combined with an XGBoost classifier, we achieved a higher average accuracy of 0.70 (variance = 0.01) and an F1-score of 0.81 (variance = 0.01). These results are based on 10 random runs, consistent with the evaluation procedure described in the original paper.

Table 1 presents the features that differ significantly between the two groups after FDR correction, highlighting associations with emotional expression, speech quality, and the duration of voiced segments. To analyze the XGBoost model results described in Section 2.2, we visualized feature im-

Feature	Non-Stressed	Stressed	p-value
Shimmer_local	0.343	-0.142	1.27×10^{-5}
Jitter_local	0.368	-0.152	4.14×10^{-4}
emotion_hap	-0.238	0.098	5.42×10^{-3}
emotion_sad	0.220	-0.091	1.08×10^{-2}
loops_L1_per_word	-0.156	0.065	1.64×10^{-2}
vader_positive	0.211	-0.087	2.54×10^{-2}
vader_compound	0.178	-0.073	2.95×10^{-2}
Number=Plur	0.186	-0.077	3.80×10^{-2}
pause_count	-0.147	0.061	4.10×10^{-2}
pause_short	-0.147	0.061	4.12×10^{-2}
F0_mean	-0.151	0.063	4.29×10^{-2}

Table 1: Mean feature values for Non-Stressed and Stressed groups (STRESSID), with corresponding p-values from two-sample t-tests ($p < 0.05$).

portance based on gain, plotted SHAP values, and examined the aggregated local explanations obtained from LIME. Figure 2 illustrates the top 10 most important features. Features related to syntax, lexical properties, and voice quality consistently emerged as the most influential across interpretability methods, with voice quality features showing the strongest overall contribution.

3.2.3 DAIC-WOZ

Binary depression classification was performed using participant speech, with subject-level acoustic and linguistic features aggregated using robust statistics and modeled with XGBoost. Classification performance was moderate (accuracy = 0.66, F1-score = 0.56, AUC-ROC = 0.63), compared with an LSTM model with an F1-score of 0.64 (Arioz et al., 2022). Statistical analysis indicated that depressed participants exhibited higher pause frequency and larger pause-to-speech ratios, reflecting reduced fluency. Linguistic differences were limited and did not remain significant after false discovery rate correction. Feature importance analyses using XGBoost, SHAP, and LIME were consistent, identifying pausing behavior, reduced pitch variability, lower vocal intensity, and increased voice quality instability as the most influential predictors.

3.2.4 ANDROIDS

Depression classification was conducted using participant-level acoustic and linguistic features aggregated across speech tasks and modeled with XGBoost. The model achieved strong performance (accuracy = 75.6%, F1-score = 77.1%, AUC-ROC = 87.6%), compared with the LSTM model reported in (Tao et al., 2023), which achieved an F1-score of 0.83. Statistical analysis revealed significant group differences in emotional expression, vocal intensity, pause behavior, and discourse structure. After false discovery rate correction, sadness-related emotion, negative sentiment, reduced intensity, and lower semantic coherence remained significant. Interpretability analyses consistently highlighted emotional polarity, voice quality measures (jitter and shimmer), vocal energy, and discourse coherence as dominant predictors.

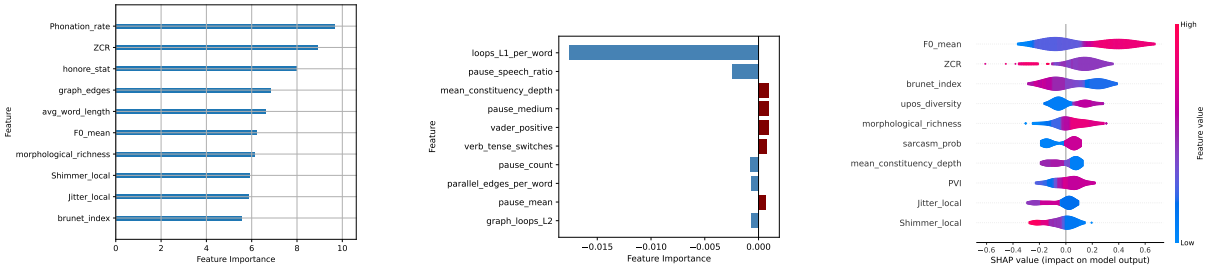
3.2.5 EATD

Depression classification was performed using multimodal acoustic and linguistic features and an XGBoost classifier. Model performance was variable (accuracy = 82.1%, F1-score = 53.9%, AUC-ROC = 73.4%), compared with a GRU model reported in (Shen et al., 2022), which achieved an F1-score of 0.71. Statistical group-level differences did not remain significant after false discovery rate correction. Despite this, feature importance analyses consistently identified reduced prosodic variability, lower articulation rate, and sadness-related emotional cues as relevant predictors.

3.2.6 Real dataset

For the REAL dataset, we converted the PHQ-9, GAD-7, and ASRS scores into binary classification tasks using clinically established thresholds: a PHQ-9 score of 15 or higher was considered indicative of at least moderate depressive symptoms, a GAD-7 score of 10 or higher indicated at least moderate anxiety symptoms, and an ASRS score of 13 or higher was considered indicative of clinically relevant ADHD symptoms. This binarization aimed to reduce variability in interpretation of self-reported questionnaire scores and to enhance consistency across participants (van Ballegooijen et al., 2016). Using an XGBoost classifier, we achieved AUC-ROC of 0.67 (variance = 0.05) for ASRS, 0.63 (variance = 0.03) for PHQ-9, and 0.59 (variance = 0.02) for GAD-7 under 4-fold, subject-independent cross-validation. All features were first aggregated to the subject level by taking the median across each participant's available audio files, and folds were created on subjects (speaker-disjoint), eliminating any train/test leakage.

Table 9 presents the features distinguishing the depressed and non-depressed groups. After applying FDR correction, the *VADER_negative* feature remained statistically significant, indicating higher levels of negative sentiment and altered emotional



(a) Stress – XGBoost (b) Stress – LIME (c) Stress – SHAP

Figure 2: Top predictive features for the STRESSID dataset derived from acoustic and linguistic descriptors.

expression in the depressed group. To further validate these findings, we compared the results with those obtained from other methodologies, as shown in Figure 3. This analysis confirmed that emotional, lexical, and syntactic features play a key role in differentiating between the two groups.

Table 3 presents the features that show statistically significant differences for the ASRS-based ADHD recognition; note that only the last four features did not remain significant after FDR correction. The most influential features appear to be those capturing repetition patterns, and the fluency of the speaker. As illustrated in Figure 3, features related to graph-based representations, verb tense variation, emotional expression, and sarcasm detection also played a key role in distinguishing individuals with high ASRS scores, indicative of ADHD. Overall, across all analyses, the frequency of verb tense shifts and graph features capturing repetition consistently emerged as the most important indicators.

Table 4 presents the most significant features identified by the t-tests for distinguishing individuals scoring above the cutoff for moderate anxiety in the GAD-7, with features associated with voice quality and emotional expression emerging as the most influential. However, after applying FDR correction, no features remained statistically significant. As shown in Figure 3, further analysis revealed that graph-based, semantic, voice quality, emotional, and lexical features played key roles in discrimination. Across all methodologies, features related to voice quality and emotional characteristics consistently proved to be the most important indicators of anxiety.

Feature	**Non-Dep**	**Dep**	p**-value**
vader_negative	0.030	0.050	1.22×10^{-4}
vader_compound	0.652	0.497	4.98×10^{-3}
content_function_ratio	1.333	1.273	9.93×10^{-3}
loops_L1_per_word	0.003	0.001	1.26×10^{-2}
pause_medium	0.295	0.102	1.91×10^{-2}
graph_loops_L1	0.302	0.148	1.77×10^{-2}
MATTR	0.693	0.704	4.78×10^{-2}
emotion_neu	0.198	0.169	4.26×10^{-2}

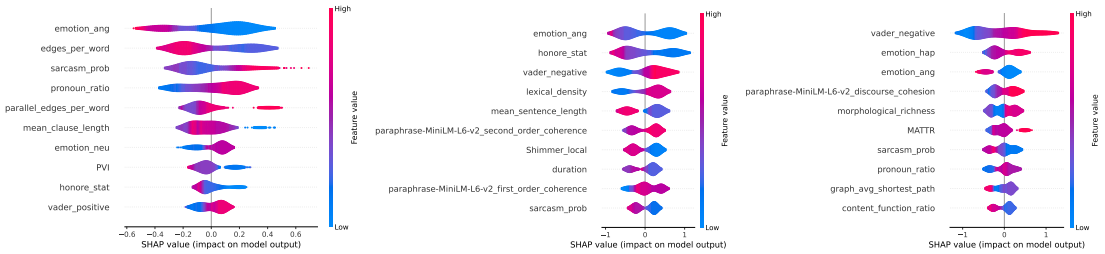
Table 2: Mean feature values for Non-Depression and Depression groups (PHQ-9) from the REAL dataset, with corresponding p-values from two-sample t-tests ($p < 0.05$).

Feature	**Non-ADHD**	**ADHD**	p**-value**
Tense=Pres	6.22	7.81	2.28×10^{-4}
graph_repeated_edges	3.12	4.48	1.75×10^{-4}
graph_diameter	8.71	8.23	3.85×10^{-4}
edges_per_word	0.93	0.92	7.15×10^{-4}
parallel_edges_per_word	0.04	0.05	1.09×10^{-3}
verb_tense_switches	8.54	10.44	2.16×10^{-3}
graph_loops_L3	1.52	1.99	4.15×10^{-3}
max_constituency_depth	13.88	14.83	5.49×10^{-3}
pronoun_ratio	0.15	0.16	5.43×10^{-3}
discourse_cohesion	0.09	0.10	7.75×10^{-3}
filler_count	116.45	136.47	7.74×10^{-3}
emotion_ang	0.06	0.04	8.68×10^{-3}
brunet_index	9.94	10.32	1.60×10^{-2}
clause_ratio	0.73	0.85	2.31×10^{-2}
honore_stat	1572.53	1522.98	4.42×10^{-2}
graph_loops_L2	0.75	0.95	4.52×10^{-2}

Table 3: Mean feature values for Non-ADHD and ADHD groups (ASRS) from the REAL dataset, with corresponding p-values from two-sample t-tests ($p < 0.05$).

Feature	**Non-Anxiety**	**Anxiety**	p**-value**
vader_negative	0.030	0.041	6.81×10^{-3}
Shimmer_local	0.111	0.103	2.17×10^{-2}

Table 4: Mean feature values for Non-Anxiety and Anxiety groups (GAD-7), with corresponding p-values from two-sample t-tests ($p < 0.05$).



(a) ASRS (b) GAD-7 (c) PHQ-9

Figure 3: SHAP explanations for top predictive features from the REAL dataset for ASRS, GAD-7, and PHQ-9.

4 Discussion

Section 3.2 suggests two key points. First, predicting psychopathology from voice is a challenging task, as also reported in (Berisha and Liss, 2024). Second, there is preliminary evidence that certain features could act as potential indicators across different conditions and clinical cases.

For anxiety and stress, Shimmer was the most prominent feature across both datasets. Shimmer measures cycle-to-cycle amplitude variation in successive glottal cycles, reflecting irregularities in vocal fold vibration. This finding is consistent with prior work (Teferra et al., 2022) and supports the potential of Shimmer as an indicator for anxiety. In our analysis, higher Shimmer values were associated with increased anxiety, in line with Jones et al. (2011), although opposite trends have also been reported (Basar et al., 2023), indicating the need for further validation before clinical use.

For the ADHD-related ASRS score, we found that the most important features were graph-based measures capturing repetition patterns (Mota et al., 2012), as well as verb-tense-related features that quantify how frequently speakers shift between tenses. This aligns with previous literature (Engelhardt et al., 2011), which reports that individuals with ADHD often exhibit higher levels of repetition and disfluency—patterns that our methodology appears capable of capturing as well.

For PHQ-9 based depression, we found that features capturing emotional content from both audio and text played an important role, along with the content-function ratio feature. The content-function ratio measures the balance between meaning-bearing words and grammatical words, indicating how information-dense or syntactically simple a text is. From this observation, we see that our system captures features reflecting both emotional expression (e.g., sadness) and syntactic richness. These findings are consistent with previous work showing that emotional and affective cues in both speech and language serve as reliable indicators of depressive states (Cummins et al., 2015). Likewise, the relevance of the content-function ratio aligns with the linguistic style framework proposed by Pennebaker and King (Pennebaker and King, 1999), where the balance between content and function words reflects cognitive style and syntactic complexity often associated with mood and mental health variations.

These findings highlight the value of an integrative, feature-centered approach to speech-based psychopathology analysis. The work combines foundation-model-based representations with traditional acoustic and linguistic features, classical statistical analysis, and interpretable machine learning across conditions and datasets. We posit that this integrated and exploratory approach is particularly well suited to clinical and interdisciplinary contexts, where transparent indicators are preferred, and may facilitate interpretable insights and hypothesis generation.

5 Conclusion

We introduced an interpretable, perceptually grounded framework for exploring speech-based correlates of psychopathology across heterogeneous datasets, languages, and labeling schemes. Using a compact set of acoustic and linguistic descriptors and transparent models with complementary explanations, we examined feature patterns (prosody/fluency, affect, voice quality, and discourse markers) that appeared to be associated with stress, depression, anxiety, and ADHD-related screening status. Future work should further investigate these candidate indicators under stronger reference standards and in longitudinal, out-of-domain settings.

6 Limitations

Despite promising within-dataset results, applying speech-based models for psychopathology prediction in real-world settings remains difficult. As highlighted by Berisha and Liss (Berisha and Liss, 2024), speech is affected by numerous confounding factors, such as fatigue, and background noise, that obscure condition-specific indicators and reduce model robustness. Differences in recording devices, linguistic and cultural backgrounds, and labeling protocols further introduce domain bias. Although some cross-setting generalization was observed, performance remained sensitive to acoustic variability and contextual shifts. Moreover, short speech samples and static features may overlook temporal cues critical for capturing symptom dynamics. Advancing clinical applicability therefore requires adaptive preprocessing, domain-invariant features, and large datasets that reflect naturalistic conditions. Finally, because several datasets use questionnaire-based cutoffs (including PHQ-8/9, GAD-7, and ASRS), the supervision signal reflects widely used, but imperfect, measurement instruments, and some errors may reflect measurement/cutoff variability rather than model behavior alone. This does not change the goal of the study (interpretability and clinical plausibility), but it does bound how literally one should read classification metrics as reflecting "true" clinical status across datasets. Additionally, neural-based features such as sarcasm detection should be interpreted cautiously, as these models have imperfect accuracy and may reproduce biases present in their training data.

7 Acknowledgment

This work was supported by PsychNow. We thank the clinical and product teams at PsychNow for their support and feedback.

References

Tuka Alhanai, Mohammad Ghassemi, and James Glass. 2018. Detecting depression with audio/text sequence modeling of interviews. In *Interspeech*, pages 1716–1720.

Nujud Aloshban, Anna Esposito, and Alessandro Vinciarelli. 2022. What you say or how you say it? depression detection through joint modeling of linguistic and acoustic aspects of speech. *Cognitive Computation*, 14(5):1585–1598.

Murray Alpert, Enrique R Pouget, and Raul R Silva. 2001. Reflections of depression in acoustic measures of the patient's speech. *Journal of affective disorders*, 66(1):59–69.

Armen C Arevian, Daniel Bone, Nikolaos Malandrakis, Victor R Martinez, Kenneth B Wells, David J Miklowitz, and Shrikanth Narayanan. 2020. Clinical state tracking in serious mental illness through computational analysis of speech. *PLoS one*, 15(1):e0225695.

Umut Arioz, Urška Smrke, Nejc Plohl, and Izidor Mlakar. 2022. Scoping review on the multimodal classification of depression and experimental study on existing multimodal models. *Diagnostics*, 12(11):2683.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Gokhan Basar, Ozlem Kaleoglu Aslan, and Mehmet Surmeli. 2023. Relationship between dysphonia and anxiety in fibromyalgia syndrome. *European Archives of Oto-Rhino-Laryngology*, 280(1):285–288.

Visar Berisha and Julie M Liss. 2024. Responsible development of clinical speech ai: Bridging the gap between clinical research and technology. *NPJ digital medicine*, 7(1):208.

Paul Boersma and David Weenink. 2021. Praat: doing phonetics by computer [Computer program]. Version 6.1.38, retrieved 2 January 2021 http://www.praat.org/.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359.

Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria. 2019. Towards multimodal sarcasm detection (an _Obviously_ perfect paper). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4619–4629, Florence, Italy. Association for Computational Linguistics.

Hava Chaptoukaev, Valeriya Strizhkova, Michele Panariello, Bianca Dalpaos, Aglind Reka, Valeria Manera, Susanne Thümmler, Esma ISMAILOVA, Nicholas W., francois bremond, Massimiliano Todisco, Maria A Zuluaga, and Laura M. Ferrari. 2023. Stressid: a multimodal dataset for stress identification. In *Advances in Neural Information Processing Systems*, volume 36, pages 29798–29811. Curran Associates, Inc.

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

Michael T Compton, Benson S Ku, Michael A Covington, Celia Metzger, and Anya Hogoboom. 2023. Lexical diversity and other linguistic measures in schizophrenia: associations with negative symptoms and neurocognitive performance. *The Journal of nervous and mental disease*, 211(8):613–620.

Cheryl M Corcoran, Facundo Carrillo, Diego Fernández-Slezak, Gillinder Bedi, Casimir Klim, Daniel C Javitt, Carrie E Bearden, and Guillermo A Cecchi. 2018. Prediction of psychosis across protocols and risk cohorts using automated language analysis. *World Psychiatry*, 17(1):67–75.

Nicholas Cummins, Alice Baird, and Bjoern W Schuller. 2018. Speech analysis for health: Current state-of-the-art and the increasing impact of deep learning. *Methods*, 151:41–54.

Nicholas Cummins, Stefan Scherer, Jarek Krajewski, Sebastian Schnieder, Julien Epps, and Thomas F Quatieri. 2015. A review of depression and suicide risk assessment using speech analysis. *Speech communication*, 71:10–49.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1*, pages 4171–4186.

Alberto Dionigi, Mirko Duradoni, and Laura Vagnoli. 2023. Understanding the association between humor and emotional distress: The role of light and dark humor in predicting depression, anxiety, and stress. *Europe's Journal of Psychology*, 19(4):358.

Jon Donnelly, Luke Moffett, Alina Jade Barnett, Hari Trivedi, Fides Schwartz, Joseph Lo, and Cynthia Rudin. 2024. Asymmirai: interpretable mammography-based deep learning model for 1–5-year breast cancer risk prediction. *Radiology*, 310(3):e232780.

Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

Paul E Engelhardt, Fernanda Ferreira, and Joel T Nigg. 2011. Language production strategies and disfluencies in multi-clause network descriptions: a study of adult attention-deficit/hyperactivity disorder. *Neuropsychology*, 25(4):442.

Eric Ettore, Philipp Müller, Jonas Hinze, Michel Benoit, Bruno Giordana, Danilo Postin, Amandine Lecomte, Hali Lindsay, P. Robert, and Alexandra König. 2022. Digital phenotyping for differential diagnosis of major depressive episode: A narrative review (preprint). *JMIR Mental Health*, 10.

Jonathan Gratch, Ron Artstein, Gale Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, David Traum, Skip Rizzo, and Louis-Philippe Morency. 2014. The distress analysis interview corpus of human and computer interviews. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3123–3128, Reykjavik, Iceland. European Language Resources Association (ELRA).

James J Gross and Hooria Jazaieri. 2014. Emotion, emotion regulation, and psychopathology: An affective science perspective. *Clinical psychological science*, 2(4):387–401.

Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A survey of methods for explaining black box models. *ACM computing surveys (CSUR)*, 51(5):1–42.

Andreas Holzinger, Georg Langs, Helmut Denk, Kurt Zatloukal, and Heimo Müller. 2019. Causability and explainability of artificial intelligence in medicine. *Wiley interdisciplinary reviews: data mining and knowledge discovery*, 9(4):e1312.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spaCy: Industrial-strength Natural Language Processing in Python.

Clayton Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the international AAAI conference on web and social media*, volume 8, pages 216–225.

Florian Hönig, Anton Batliner, Elmar Nöth, Sebastian Schnieder, and Jarek Krajewski. 2014. Automatic modelling of depressed speech: relevant features and relevance of gender. In *Interspeech 2014*, pages 1248–1252.

Yannick Jadoul, Bill Thompson, and Bart de Boer. 2018. Introducing Parselmouth: A Python interface to Praat. *Journal of Phonetics*, 71:1–15.

Yishan Jiao, Visar Berisha, and Julie Liss. 2017. Interpretable phonological features for clinical applications. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 5045–5049. IEEE.

Mark Jones, Flora Anagnostou, and Jo Verhoeven. 2011. The vocal expression of emotion: An acoustic analysis of anxiety. In *ICPhS*, pages 982–985.

Daniel Korzekwa, Roberto Barra-Chicote, Bozena Kostek, Thomas Drugman, and Mateusz Lajszczak. 2019. Interpretable Deep Learning Model for the Detection and Reconstruction of Dysarthric Speech. In *Interspeech 2019*, pages 3890–3894.

Roman Kotov, Robert Krueger, David Watson, Thomas Achenbach, Robert Althoff, R. Bagby, Timothy Brown, William Carpenter, Avshalom Caspi, Lee Clark, Nicholas Eaton, Miriam Forbes, Kelsie Forbush, David Goldberg, Deborah Hasin, Steven Hyman, Masha Ivanova, Donald Lynam, Kristian Markon, and Mark Zimmerman. 2017. The hierarchical taxonomy of psychopathology (hitop): A dimensional alternative to traditional nosologies. *Journal of Abnormal Psychology*, 126:454–477.

Emma CL Leschly, Oliver Roesler, Michael Neumann, Jackson Liscombe, Abhishek Hosamath, Lakshmi Arbatti, Line H Clemmensen, Melanie Ganz, and Vikram Ramanarayanan. 2025. An exploration of interpretable deep learning models for the assessment of mild cognitive impairment. In *Proc. Interspeech 2025*, pages 271–275.

Daniel M Low, Kate H Bentley, and Satrajit S Ghosh. 2020. Automated assessment of psychiatric disorders using speech: A systematic review. *Laryngoscope investigative otolaryngology*, 5(1):96–116.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Felix Menne, Felix Dörr, Julia Schräder, Johannes Tröger, Ute Habel, Alexandra König, and Lisa Wagels. 2024. The voice of depression: speech features as biomarkers for major depressive disorder. *BMC psychiatry*, 24(1):794.

Natalia B. Mota, Nathaly A. P. Vasconcelos, Nilza Lemos, Ana C. Pieretti, Osame Kinouchi, Guillermo A. Cecchi, Mauro Copelli, and Sidarta Ribeiro. 2012. Speech graphs provide a quantitative measure of thought disorder in psychosis. *PLOS ONE*, 7(4):e34928.

Michael Neumann, Hardik Kothare, Beverly Insel, Anzalee Khan, Danyah Nadim, Jean-Pierre Lindenmayer, and Vikram Ramanarayanan. 2025. Multimodal speech, language and orofacial analysis for remote assessment of positive, negative and cognitive symptoms in schizophrenia. In *Proc. Interspeech 2025*, pages 5703–5707.

Si-Ioi Ng, Lingfeng Xu, Ingo Siegert, Nicholas Cummins, Nina R Benway, Julie Liss, and Visar Berisha. 2024. A tutorial on clinical speech ai development: From data collection to model validation. *arXiv preprint arXiv:2410.21640*.

Stavros Ntalampiras. 2025. Interpretable probabilistic identification of depression in speech. *Sensors*, 25(4):1270.

James W Pennebaker and Laura A King. 1999. Linguistic styles: language use as an individual difference. *Journal of personality and social psychology*, 77(6):1296.

James W Pennebaker, Matthias R Mehl, and Kate G Niederhoffer. 2003. Psychological aspects of natural language use: Our words, our selves. *Annual review of psychology*, 54(1):547–577.

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

Gowtham Premananth, Philip Resnik, Sonia Bansal, Deanna L. Kelly, and Carol Espy-Wilson. 2025. Multimodal Biomarkers for Schizophrenia: Towards Individual Symptom Severity Estimation. In *Interspeech 2025*, pages 3065–3069.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th annual meeting of the association for computational linguistics: system demonstrations*, pages 101–108.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 3982–3992.

Rachid Riad, Martin Denais, Marc de Gennes, Adrien Lesage, Vincent Oustric, Xuan Nga Cao, Stéphane Mouchabac, and Alexis Bourla. 2024. Automated speech analysis for risk detection of depression, anxiety, insomnia, and fatigue: algorithm development and validation study. *Journal of Medical Internet Research*, 26:e58572.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Ying Shen, Huiyu Yang, and Lin Lin. 2022. Automatic depression detection: An emotional audio-textual corpus and a gru/bilstm-based model. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6247–6251. IEEE.

Anke R Sonnenschein, Stefan G Hofmann, Tobias Ziegelmayer, and Wolfgang Lutz. 2018. Linguistic analysis of patients with mood and anxiety disorders during cognitive behavioral therapy. *Cognitive behaviour therapy*, 47(4):315–327.

Jee Eun Sung, Sujin Choi, Bora Eom, Jae Keun Yoo, and Jee Hyang Jeong. 2020. Syntactic complexity as a linguistic marker to differentiate mild cognitive impairment from normal aging. *Journal of Speech, Language, and Hearing Research*, 63(5):1416–1429.

Fuxiang Tao, Anna Esposito, and Alessandro Vinciarelli. 2023. The Androids Corpus: A New Publicly Available Benchmark for Speech Based Depression Detection. In *Interspeech 2023*, pages 4149–4153.

Mashrura Tasnim, Malikeh Ehghaghi, Brian Diep, and Jekaterina Novikova. 2022. DEPAC: a corpus for depression and anxiety detection from speech. In *Proceedings of the Eighth Workshop on Computational Linguistics and Clinical Psychology*, pages 1–16, Seattle, USA. Association for Computational Linguistics.

Bazen Gashaw Teferra, Sophie Borwein, Danielle D DeSouza, William Simpson, Ludovic Rheault, and Jonathan Rose. 2022. Acoustic and linguistic features of impromptu speech and their association with anxiety: validation study. *JMIR mental health*, 9(7):e36828.

Ming Tu, Visar Berisha, and Julie Liss. 2017. Interpretable objective assessment of dysarthric speech based on deep neural networks. In *Interspeech*, pages 1849–1853.

Wouter van Ballegooijen, Heleen Riper, Pim Cuijpers, Patricia van Oppen, and Johannes H Smit. 2016. Validation of online psychometric instruments for common mental health disorders: a systematic review. *BMC psychiatry*, 16(1):45.

Rohit Voleti, Julie M Liss, and Visar Berisha. 2019. A review of automated speech and language features for assessment of cognitive and thought disorders. *IEEE journal of selected topics in signal processing*, 14(2):282–298.

Theo Vos, {Stephen S} Lim, Cristiana Abbafati, {Kaja M} Abbas, Mohammad Abbasi, Mitra Abbasifard, Mohsen Abbasi-Kangevari, Hedayat Abbastabar, Foad Abd-Allah, Ahmed Abdelalim, Mohammad Abdollahi, Ibrahim Abdollahpour, Hassan Abolhassani, Victor Aboyans, {Elissa M} Abrams, {Lucas Guimarães} Abreu, {Michael R M} Abrigo, {Laith Jamal} Abu-Raddad, {Abdelrahman I} Abushouk, and 1011 others. 2020. Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: a systematic analysis for the global burden of disease study 2019. *The Lancet*, 396(10258):1204–1222.

Lingfeng Xu, Julie Liss, and Visar Berisha. 2023. Dysarthria detection based on a deep learning model with a clinically-interpretable layer. *JASA Express Letters*, 3(1).

Shu-Wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y. Lin, Andy T. Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, Tzu-Hsien Huang, Wei-Cheng Tseng, Ko tik Lee, Da-Rong Liu, Zili Huang, Shuyan Dong, Shang-Wen Li, Shinji Watanabe, Abdelrahman Mohamed, and Hung yi Lee. 2021. SUPERB: Speech Processing Universal PERformance Benchmark. In *Interspeech 2021*, pages 1194–1198.

A Demographic Data

This section summarizes the demographic and clinical score distributions across the evaluated datasets. Where available, standardized psychometric instruments are used to characterize symptom severity and population variability.

For the REAL dataset, Figure 4 presents distributions of age and self-reported clinical scales. Participant ages span a broad range, with most individuals between their 20s and 50s. ASRS scores show moderate dispersion with a noticeable concentration at higher values, indicating a substantial proportion of participants exhibiting elevated ADHD-related traits. GAD-7 scores are broadly distributed, suggesting balanced representation across anxiety severity levels, while PHQ-9 scores span the full clinical range, with many participants in the moderate to severe depression categories.

For the DAIC-WOZ corpus, Figure 5 illustrates the distribution of PHQ-8 scores obtained during semi-structured clinical interviews. The distribution reflects substantial variability in depressive symptom severity, supporting its use as a benchmark dataset for depression detection from speech.

Figure 6 shows the distribution of SDS scores for the EATD dataset. Most participants fall below the clinical threshold, with a smaller subset exhibiting elevated depression scores, consistent with the dataset's class imbalance and student-based recruitment.

B Feature Distributions

To compare the feature distributions between the *Stress* and REAL datasets, we visualize the normalized histograms for each extracted feature (see Figure 7). Red denotes the Stress dataset, while blue corresponds to REAL. Overall, the majority of features display similar distributional patterns between the *Stress* and REAL datasets. This consistency supports the validity and reliability of our feature extraction procedure, suggesting that the computed features capture comparable underlying characteristics across recording conditions.

C Partial Dependence Plots

Partial Dependence Plots (PDPs) illustrate the marginal effect of individual features on model predictions while averaging over all other variables. By doing so, they reveal whether a given feature has a positive, negative, or non-linear influence on the target outcome. PDPs are particularly informative for complex, non-linear models such as ensemble or deep learning architectures, as they provide interpretable insight into learned feature-outcome relationships.

Figure 8 presents the PDPs for a range of speech-derived features across the *Real* dataset models (ASRS, GAD-7, and PHQ-9). These plots highlight how prosodic, lexical, and semantic attributes contribute to predicting mental health symptom severity.

Figure 9 focuses on the STRESSID model, illustrating the partial dependencies for key linguistic and acoustic predictors. The monotonic and threshold-based trends observed, such as the influence of speech rate, pause duration, and pitch variability, suggest interpretable and physiologically plausible mappings between vocal behavior and stress-related outcomes.

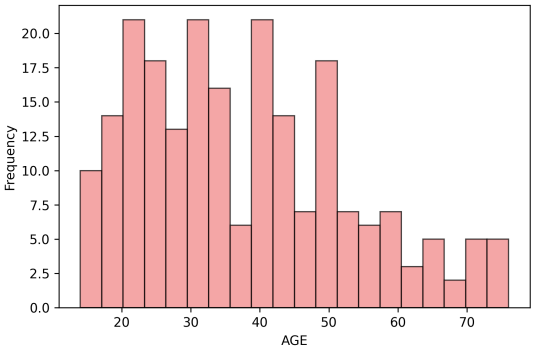
The Partial Dependence Plots for the DAIC-WOZ dataset (Figure 10) show that increased pause duration and reduced pitch variability are associated with higher depression probability. Elevated jitter and shimmer further indicate reduced phonatory stability in depressed speech. Linguistic features such as shorter word length and lower lexical diversity exhibit monotonic trends, reflecting cognitive and expressive slowing.

For the EATD corpus, PDPs (Figure 12) reveal strong non-linear effects, with higher articulation rate and wider pitch range associated with lower depression likelihood. Increased sad emotion activation sharply increases predicted depression probability. Linguistic simplification beyond threshold levels further contributes to elevated risk.

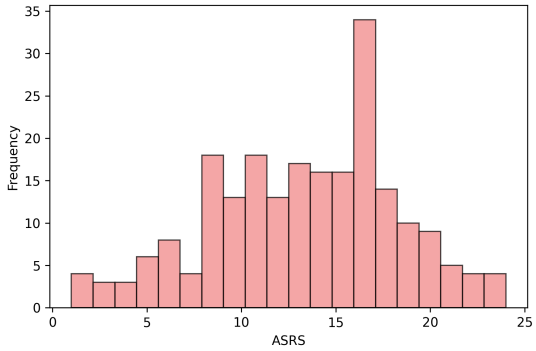
In the ANDROIDS CORPUS, PDPs (Figure 11) indicate that reduced vocal intensity, increased shimmer, and longer pauses correspond to higher depression probability. Emotional polarity features show strong monotonic relationships, with higher sadness and lower positive sentiment increasing risk. Disrupted discourse coherence further amplifies depression predictions across tasks.

D Feature Descriptions

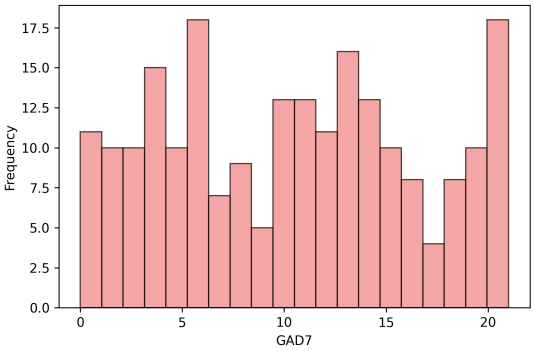
To ensure interpretability and transparency in our modeling pipeline, we provide a detailed description of all extracted acoustic and linguistic features used in our analyses, see Table 5. Each feature is annotated with its extraction method, type, subtype, and psycholinguistic interpretation. These features



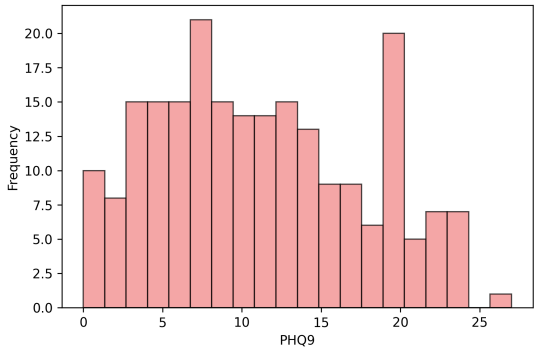
Age



ASRS



GAD-7



PHQ-9

Figure 4: Distributions of demographic and clinical variables in the REAL dataset.

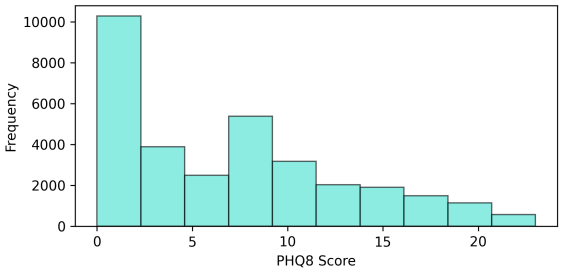


Figure 5: Distribution of PHQ-8 depression scores in the DAIC-WOZ dataset.

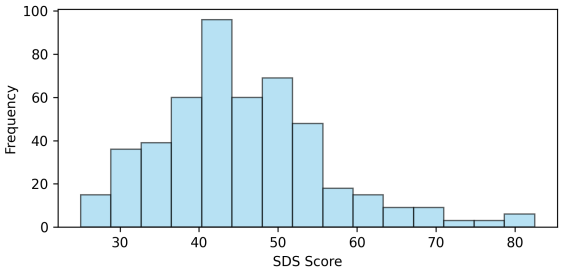


Figure 6: Distribution of SDS depression scores in the EATD dataset.

encompass prosodic, voice quality, lexical, syntactic, semantic, and psycholinguistic dimensions, thereby capturing both low-level acoustic cues and higher-level cognitive-linguistic markers of mental state.

E Feature Correlation

Figure 13 presents correlation matrices of the extracted acoustic and linguistic features across all evaluated datasets. The matrices illustrate pairwise linear relationships among features, revealing structured intra-group correlations and dataset-specific differences in feature dependence. Variations in correlation strength across datasets reflect differences in recording conditions, task design, and population characteristics, and highlight potential feature redundancy addressed by the modeling approach.

F Additional Results and Cross-Dataset Analysis

This section presents additional results obtained using the proposed analytical framework across multiple clinical and real-world speech datasets,

Feature	Type	Subtype	Calculation / Definition
ZCR	Acoustic	Prosodic	Zero-crossing rate of the normalized voiced waveform.
F0_mean	Acoustic	Prosodic	Mean pitch (Hz) from Praat pitch analysis.
F0_range	Acoustic	Prosodic	Maximum pitch minus minimum pitch.
F0_var	Acoustic	Prosodic	Variance of pitch values across voiced frames.
F0_std	Acoustic	Prosodic	Standard deviation of F0 values.
Intensity_mean	Acoustic	Prosodic	Mean amplitude of the waveform.
Intensity_std	Acoustic	Prosodic	Standard deviation of intensity values.
Jitter_local	Acoustic	Voice Quality	Mean absolute F0 period difference.
Shimmer_local	Acoustic	Voice Quality	Mean absolute amplitude difference.
HNR	Acoustic	Voice Quality	Harmonics-to-noise ratio (Praat Harmonicity).
PVI	Acoustic	Voice Quality	Pairwise variability index across voiced intervals.
filler_count	Linguistic	Voice Quality	Count of filler tokens such as "um", "uh", "erm".
duration	Acoustic	Prosodic	Total length of the audio file (seconds).
Phonation_rate	Acoustic	Prosodic	Number of voiced frames divided by total duration.
pause_count	Acoustic	Prosodic	Count of silent segments below amplitude threshold.
pause_short	Acoustic	Prosodic	Pauses < 1 s.
pause_medium	Acoustic	Prosodic	Pauses $1-2$ s.
pause_long	Acoustic	Prosodic	Pauses > 2 s.
pause_mean	Acoustic	Prosodic	Mean duration of all pauses (s).
pause_speech_ratio	Acoustic	Prosodic	Total pause time divided by total duration.
articulation_rate	Acoustic	Prosodic	Voiced frames / (duration $-$ total pause time).
speech_entropy	Acoustic	Prosodic	Shannon entropy of voiced amplitudes.
emotion_neu	Acoustic	Psycholinguistic	Neutral emotion probability (HuBERT model).
emotion_hap	Acoustic	Psycholinguistic	Happiness probability (HuBERT model).
emotion_ang	Acoustic	Psycholinguistic	Anger probability (HuBERT model).
emotion_sad	Acoustic	Psycholinguistic	Sadness probability (HuBERT model).
word_count	Linguistic	Lexical	Total number of words in transcript.
sentence_count	Linguistic	Lexical	Total number of sentences in transcript.
type_token_ratio	Linguistic	Lexical	Ratio of unique words to total words.
MATTR	Linguistic	Lexical	Moving-average type-token ratio (window=50).
brunet_index	Linguistic	Lexical	$N^{(V^{-0.165})}$ (lexical richness).
honore_stat	Linguistic	Lexical	$100 \log N/(1 - H/V)$ (lexical richness measure).
lexical_density	Linguistic	Lexical	Ratio of content words to total words.
idea_density	Linguistic	Lexical	Ratio of idea-bearing words to total words.
content_function_ratio	Linguistic	Lexical	Ratio of content to function words.
pronoun_ratio	Linguistic	Lexical	Pronouns divided by total words.
Tense_Past	Linguistic	Lexical	Count of past tense verbs.
Tense_Pres	Linguistic	Lexical	Count of present tense verbs.
Voice_Pass	Linguistic	Lexical	Count of passive voice verbs.
Number_Plur	Linguistic	Lexical	Count of plural nouns or verbs.
lemma_ttr	Linguistic	Lexical	Unique lemmas / total lemmas.
upos_diversity	Linguistic	Lexical	Unique UPOS tags / total tokens.
morphological_richness	Linguistic	Lexical	Unique morph features / total words.
propositional_density	Linguistic	Lexical	Ratio of verbs, adj, adv, prep, conj to total words.
mean_sentence_length	Linguistic	Syntactic	Mean words per sentence.
mean_clause_length	Linguistic	Syntactic	Mean number of tokens per clause.
syntactic_depth_mean	Linguistic	Syntactic	Mean dependency tree depth.
syntactic_depth_max	Linguistic	Syntactic	Maximum dependency tree depth.
clause_ratio	Linguistic	Syntactic	Clauses per sentence.
verb_tense_switches	Linguistic	Syntactic	Count of verb tense changes.
verb_tense_switch_ratio	Linguistic	Syntactic	Tense switches / total verbs.
syntactic_embedding_depth	Linguistic	Syntactic	Maximum dependency embedding depth.
passive_voice_ratio	Linguistic	Syntactic	Passive voice verbs / total verbs.
graph_nodes	Linguistic	Syntactic	Number of unique nodes in the speech graph.
graph_edges	Linguistic	Syntactic	Number of total edges in the speech graph.
graph_repeated_edges	Linguistic	Syntactic	Count of edges appearing multiple times.
graph_largest_scc	Linguistic	Syntactic	Size of largest strongly connected component.
graph_density	Linguistic	Syntactic	Ratio of actual to maximum possible edges.
graph_loops_L1	Linguistic	Syntactic	Number of self-loops (1-node cycles).
graph_loops_L2	Linguistic	Syntactic	Number of 2-node cycles.
graph_loops_L3	Linguistic	Syntactic	Number of 3-node cycles.
graph_avg_total_degree	Linguistic	Syntactic	Average in+out degree across nodes.
graph_diameter	Linguistic	Syntactic	Longest shortest path in undirected graph.
graph_avg_shortest_path	Linguistic	Syntactic	Mean shortest path length in graph.
nodes_per_word	Linguistic	Syntactic	graph_nodes / total_words.
edges_per_word	Linguistic	Syntactic	graph_edges / total_words.
atd_per_word	Linguistic	Syntactic	graph_avg_total_degree / total_words.
parallel_edges_per_word	Linguistic	Syntactic	graph_repeated_edges / total_words.
loops_L1_per_word	Linguistic	Syntactic	graph_loops_L1 / total_words.
loops_L2_per_word	Linguistic	Syntactic	graph_loops_L2 / total_words.
loops_L3_per_word	Linguistic	Syntactic	graph_loops_L3 / total_words.
mean_constituency_depth	Linguistic	Syntactic	Mean depth of Stanza constituency tree.
max_constituency_depth	Linguistic	Syntactic	Maximum depth of Stanza constituency tree.
first_order_coherence	Linguistic	Semantic	Mean cosine simmilarity between sentences.
second_order_coherence	Linguistic	Semantic	Average similarity between sentences separated by one.
discourse_cohesion	Linguistic	Semantic	Mean token overlap between consecutive sentences.
sentence_repetition_ratio	Linguistic	Semantic	Exact repeated sentences / total sentences.
vader_negative	Linguistic	Psycholinguistic	VADER negative sentiment proportion.
vader_neutral	Linguistic	Psycholinguistic	VADER neutral sentiment proportion.
vader_positive	Linguistic	Psycholinguistic	VADER positive sentiment proportion.
vader_compound	Linguistic	Psycholinguistic	Composite sentiment score in range $[-1, 1]$.
sarcasm_prob	Bimodal	Psycholinguistic	Probability of sarcasm (BERT + Wav2Vec classifier).

Table 5: List of 82 acoustic and linguistic features with their computational definitions and grouping by type and subtype.

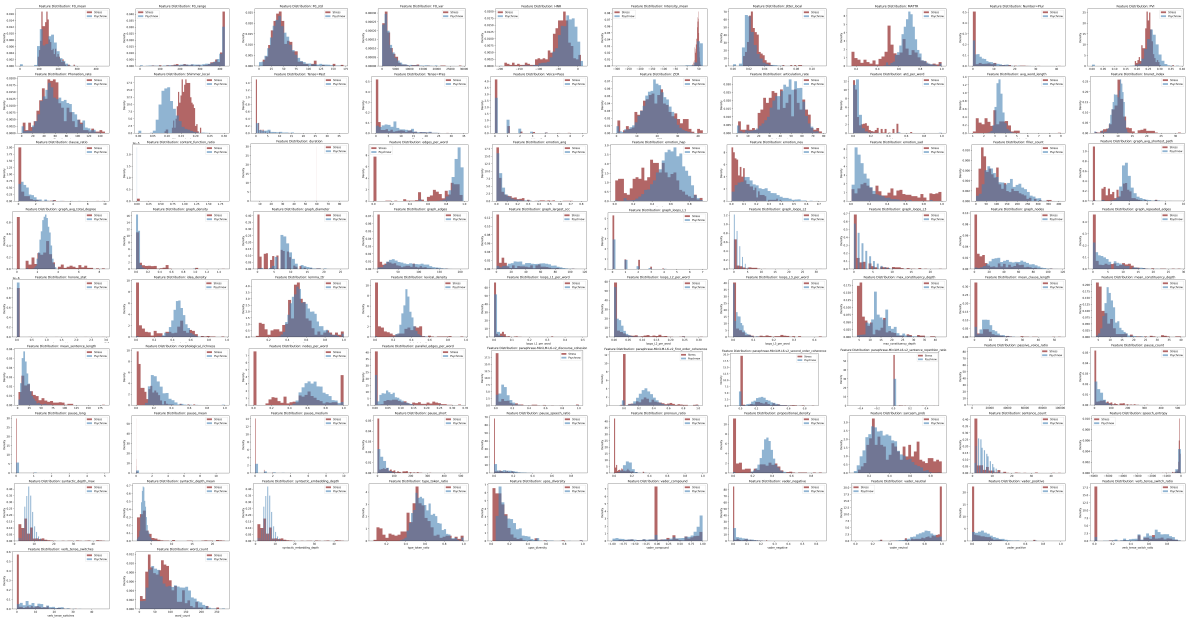


Figure 7: Comparison of feature value distributions between the *Stress* (maroon) and REAL (steelblue) datasets across 82 extracted features.

including STRESSID, DAIC-WOZ, ANDROIDS, EATD, and the REAL dataset. We report both interpretable model explanations and statistical group comparisons to provide complementary perspectives on feature relevance. Figures illustrate feature importance rankings derived from XGBoost built-in importance, LIME, and SHAP, while accompanying tables summarize statistically significant differences between clinical and non-clinical groups based on two-sample t-tests. Across datasets and mental health conditions, certain patterns can be observed, suggesting a potential role for prosodic variability, emotional expression, lexical richness, and graph-based linguistic structure as indicators of stress, depression, anxiety, and attentional symptoms.

Feature	**Non-Depressed**	**Depressed**	**p-value**
sarcasm_prob	0.465	0.431	2.44×10^{-3}
pause_short	0.357	0.768	4.13×10^{-3}
pause_speech_ratio	0.002	0.006	6.67×10^{-3}
avg_word_length	3.540	3.490	5.28×10^{-2}

Table 6: Mean feature values for Non-Depressed and Depressed groups (DAIC-WOZ), with corresponding p-values from two-sample t-tests ($p < 0.05$). Features were standardized (z-scored) prior to analysis.

Feature	**Non-Depressed**	**Depressed**	**p-value**
emotion_sad	0.099	0.182	2.83×10^{-5}
Intensity_mean	58.360	54.726	2.73×10^{-5}
sentiment_negative	0.519	0.693	6.49×10^{-5}
sentiment_positive	0.469	0.300	9.91×10^{-5}
second_order_coherence	0.213	0.313	1.08×10^{-3}
Jitter_local	0.021	0.024	1.26×10^{-2}
emotion_neu	0.199	0.156	1.96×10^{-2}
Shimmer_local	0.110	0.122	1.84×10^{-2}
brunet_index	8.852	9.740	1.70×10^{-2}
duration	54.971	72.298	3.56×10^{-2}
Phonation_rate	108.486	94.456	3.73×10^{-2}
pause_count	1.074	3.594	4.45×10^{-2}
pause_short	1.074	3.594	4.45×10^{-2}
pause_mean	0.0013	0.0038	3.63×10^{-2}
pause_speech_ratio	0.00035	0.00119	4.03×10^{-2}
type_token_ratio	0.698	0.646	2.68×10^{-2}
lemma_ttr	0.628	0.575	4.64×10^{-2}
sentence_count	5.212	6.536	3.19×10^{-2}
first_order_coherence	0.309	0.375	3.94×10^{-2}
upos_diversity	0.276	0.210	2.44×10^{-2}
emotion_hap	0.662	0.611	4.94×10^{-2}
MATTR	0.796	0.775	5.32×10^{-2}

Table 7: Mean acoustic and linguistic feature values for Non-Depressed and Depressed participants in the ANDROIDS corpus, with corresponding p-values from two-sample t-tests ($p < 0.05$ shown).

Feature	**Non-Depressed**	**Depressed**	**p-value**
emotion_neu	0.112	-0.493	8.58×10^{-4}
emotion_sad	-0.128	0.562	1.49×10^{-2}
F0_mean	-0.102	0.448	1.86×10^{-2}
passive_voice_ratio	0.046	-0.201	2.82×10^{-2}
pause_count	0.052	-0.230	4.78×10^{-2}

Table 8: Mean acoustic and linguistic feature values for Non-Depressed and Depressed participants in the EATD corpus, with corresponding p-values from two-sample t-tests ($p < 0.05$ shown).

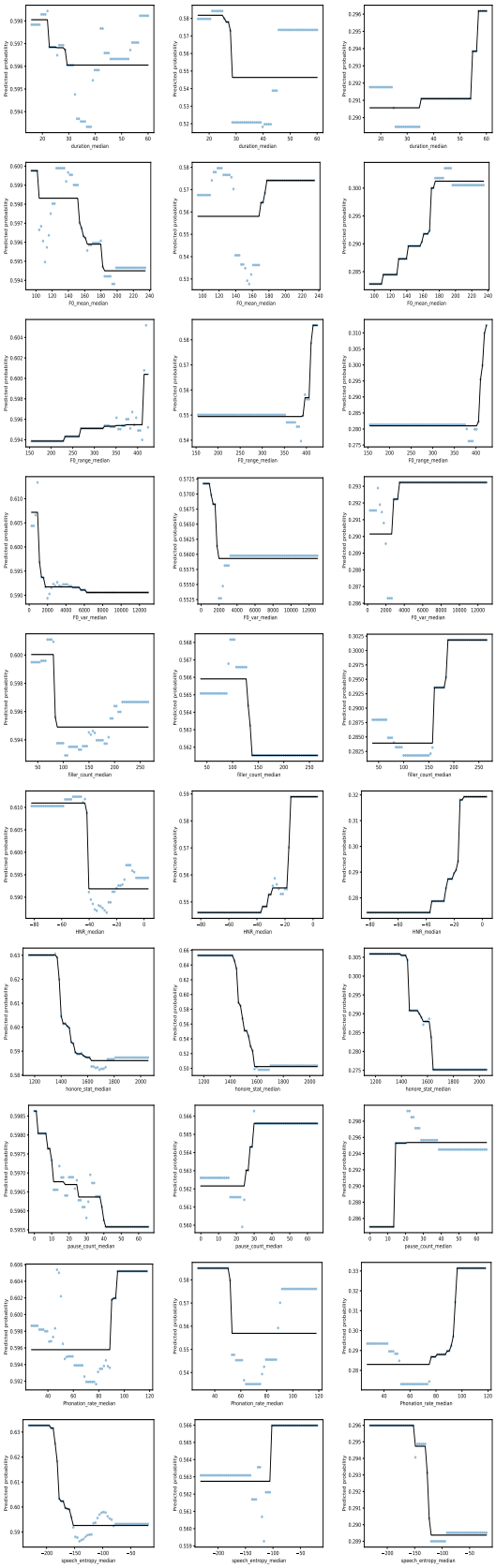


Figure 8: Partial Dependence Plots for speech-derived features across the REAL model predictions for ASRS (first column), GAD-7 (second column), and PHQ-9 (third column).

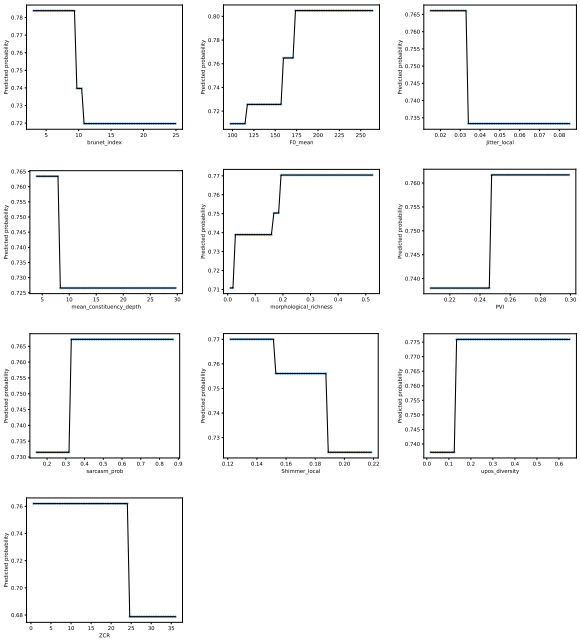


Figure 9: Partial Dependence Plots for key linguistic and acoustic features in the STRESSID model. Each curve shows the monotonic relationship between the feature value and model prediction.

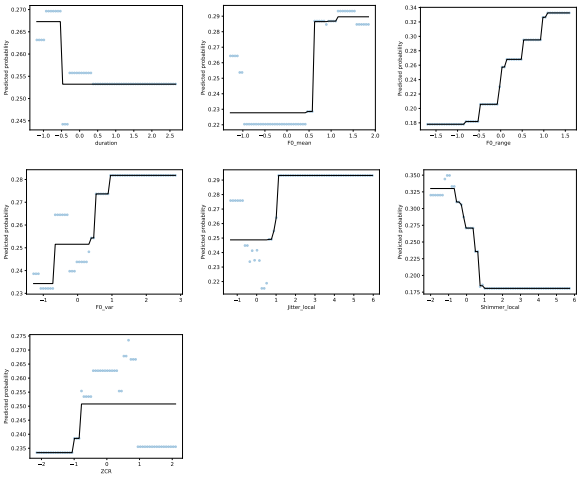


Figure 10: Partial Dependence Plots for key acoustic features in the DAIC-WOZ model.

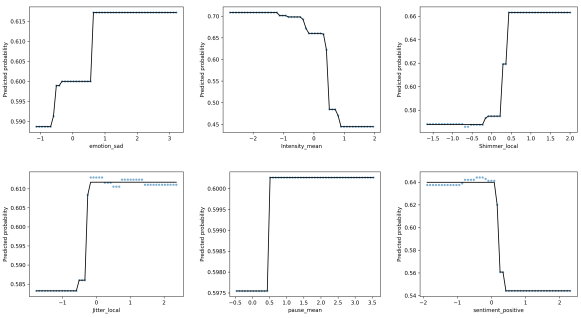


Figure 11: Partial Dependence Plots for key acoustic, emotional, and linguistic features in the ANDROIDS model.

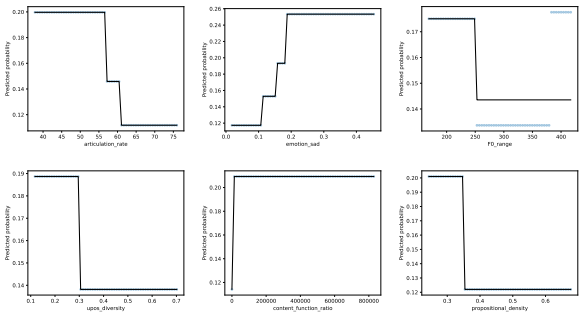


Figure 12: Partial Dependence Plots for key acoustic and linguistic features in the EATD model.

Feature	Non-Depressed	Depressed	p-value
vader_negative	0.030	0.050	1.22×10^{-4}
vader_compound	0.652	0.497	4.98×10^{-3}
content_function_ratio	1.333	1.273	9.93×10^{-3}
loops_L1_per_word	0.003	0.001	1.26×10^{-2}
pause_medium	0.295	0.102	1.91×10^{-2}
graph_loops_L1	0.302	0.148	1.77×10^{-2}
MATTR	0.693	0.704	4.78×10^{-2}
emotion_neu	0.198	0.169	4.26×10^{-2}

Table 9: Mean feature values for Non-Depression and Depression groups (PHQ-9) from the REAL dataset, with corresponding p-values from two-sample t-tests ($p < 0.05$).

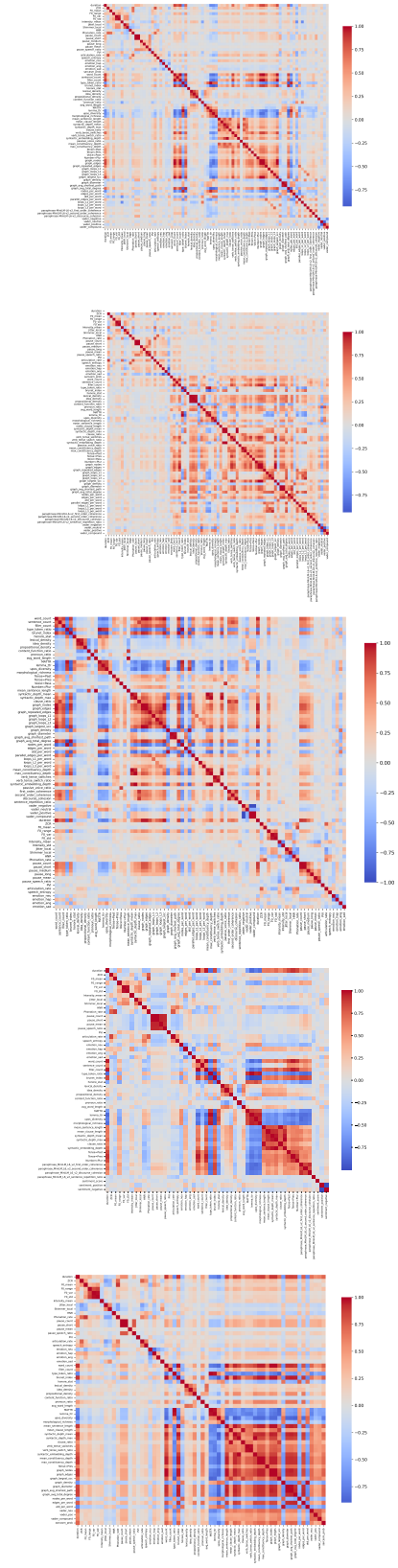
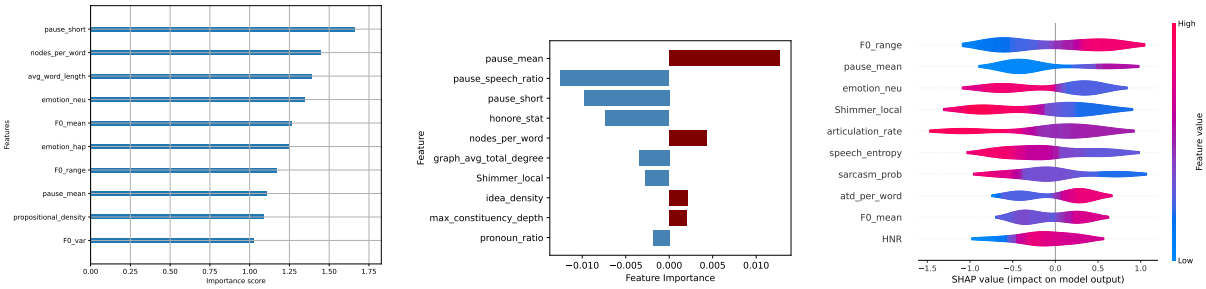
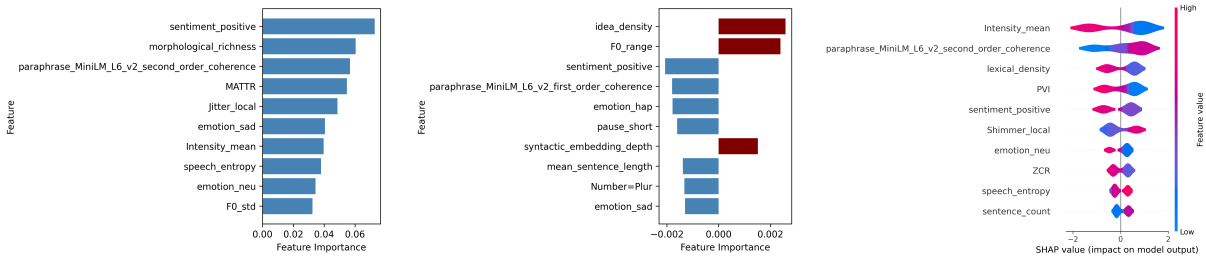


Figure 13: Correlation matrices of the extracted acoustic and linguistic features across datasets. From top to bottom, panels correspond to REAL, STRESSID, EATD, ANDROIDS, and DAIC-WOZ, respectively. Each figure shows pairwise linear correlations among features.



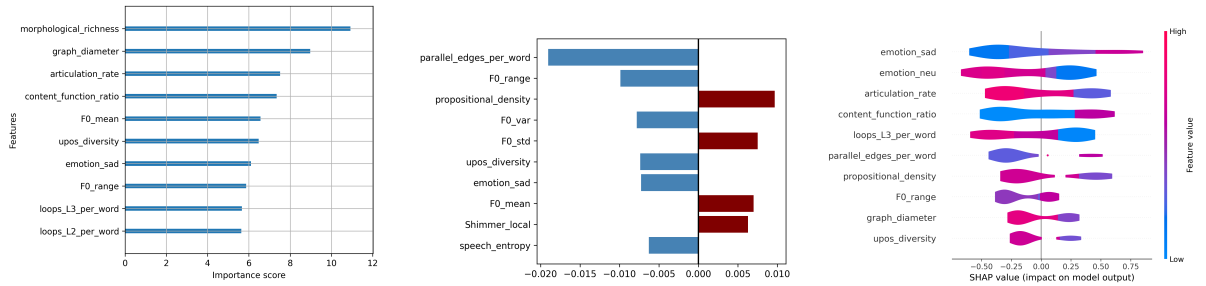
(a) DAIC-WOZ – XGBoost (b) DAIC-WOZ – LIME (c) DAIC-WOZ – SHAP

Figure 14: Top predictive features for the DAIC-WOZ dataset derived from acoustic and linguistic descriptors.



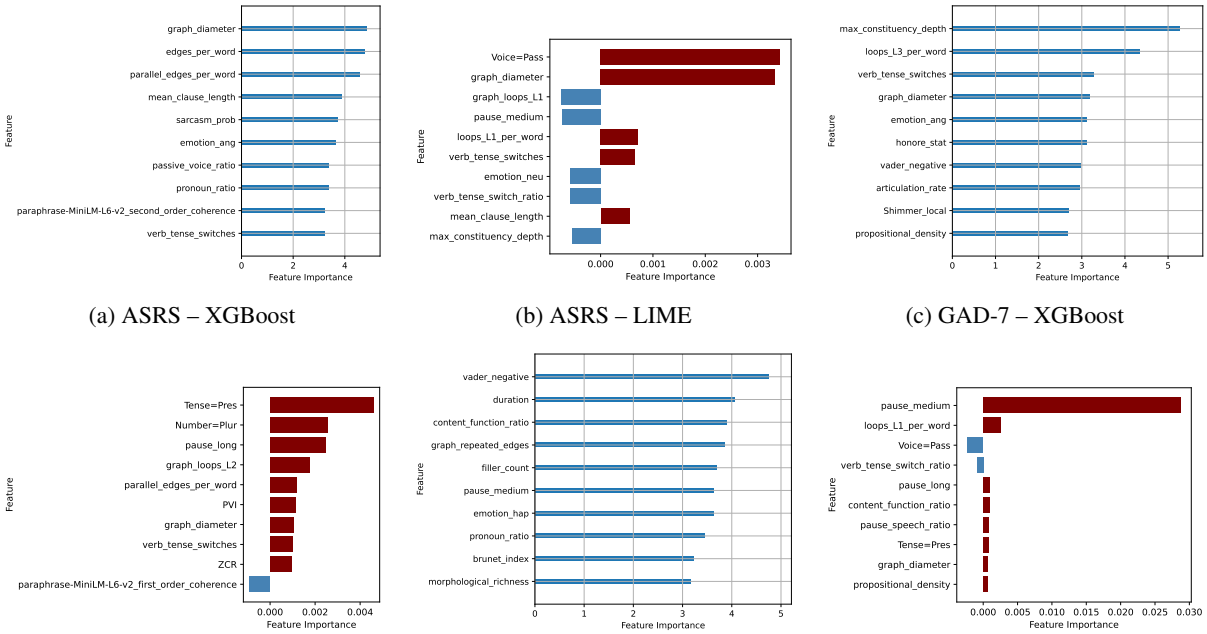
(a) Androids – XGBoost (b) Androids – LIME (c) Androids – SHAP

Figure 15: Top predictive features for the ANDROIDS CORPUS dataset derived from acoustic and linguistic descriptors.



(a) EATD – XGBoost (b) EATD – LIME (c) EATD – SHAP

Figure 16: Top predictive features for the EATD dataset derived from acoustic and linguistic descriptors.



(a) ASRS – XGBoost (b) ASRS – LIME (c) GAD-7 – XGBoost

(d) GAD-7 – LIME (e) PHQ-9 – XGBoost (f) PHQ-9 – LIME

Figure 17: Top predictive features from the REAL dataset for ASRS, GAD-7, and PHQ-9 classification tasks using XGBoost built-in importance and LIME.