

NLP and Machine Learning Approaches for Screening Depression and Anxiety via Linguistic and Acoustic Features

A Multimodal Computational Framework for Non-Invasive Mental Health Screening

Sidra Rafiq¹, Ahsan Raza², Name Hafiz Muzzammil Nasir³, Fareeha Syed⁴

¹ COMSATS University Islamabad, Email: sidrarafiq957@gmail.com

² Department of English, Government College University Faislaabad (GCUF), Pakistan.

Email: ahsanrazaofficial@gmail.com

³ Department English, University of Education Lahore vehari campus.

Email: muzzammilnasir19@gmail.com

⁴ Department of Education and Psychology, University of Aveiro, Email: fareeha.syed@ua.pt

DOI: <https://doi.org/10.70670/sra.v4i1.2441>

Abstract

Depression and anxiety disorders are also one of the major causes of disability in the world, but there is a high level of underdiagnosis because of access to clinical resources, stigma and subjectivity of traditional screening tools. This paper introduces a multimodal computational framework for short, semi-structured speech data using natural language processing (NLP) and acoustic signal analysis to detect depressed and anxious symptomatology. A collection of 460 recorded clinical interviews is used to extract a rich set of linguistic markers (lexical, syntactic, and psycholinguistic, based on dictionaries like LIWC) and acoustic markers (prosodic, spectral, and voice-quality) from the text. A late-fusion of six different machine learning architectures: Logistic Regression, Support Vector Machines, Random Forest, XGBoost, a Bidirectional LSTM, and a fine-tuned BERT encoder are benchmarked separately and in combination. The fusion model performs best in the case of depression screening (accuracy = 87.6%, F1 = 86.9%, AUC-ROC = 0.92) and anxiety screening (accuracy = 84.3%, F1 = 83.1%, AUC-ROC = 0.89) compared with unimodal baselines. The feature-importance analysis shows that pause ratio, pitch variability, speech rate, and negative-emotion word frequency are the most discriminative features for both conditions. In addition to raw performance, we measure the incremental contribution of each feature family using systematic ablation, we study the performance of each subset of the demograph, we identify potential fairness gaps, and we measure inference time to gauge feasibility for near-real-time screening. The clinical implications, generalizability and fairness limitations and ethical safeguards before deployment of such systems as screening aids, rather than diagnostic replacements, are discussed.

Keywords: *Depression Detection; Anxiety Screening; Natural Language Processing; Speech Acoustic Features; Multimodal Fusion; Mental Health Informatics; Machine Learning; Deep Learning*

Introduction

Major depression disorder (MDD) and generalized anxiety (GAD) are ranked as among the biggest causes of years lived with disability (YLD) and affect several hundred million people worldwide. Screening is not widely used in primary care and community settings, despite the existence of validated forms or questionnaires like the PHQ-9 and GAD-7, due to a variety of factors, including time constraints, shortage of clinicians, and patients' lack of willingness to self-report symptoms directly. This has sparked an increasing interest in studying mental health status indicators derived from unobtrusive and non-invasive behavior (such as verbal or written) and without burden.

The public-health burden due to underdiagnosis is significant. Depression and anxiety are linked to increased risk of chronic disease, work impairment, breakdown of relationships and suicide and the time between onset of symptoms and the start of treatment is often years long. Traditional screening pathways rely on patients actively seeking care and honestly reporting symptoms of concern during an office visit, a practice that is less likely to be observed among groups experiencing cultural stigma or language barriers, or among those who have had negative experiences in mental health care. Passive behavioural sensing provides a complementary pathway: speech is ongoing and spontaneous, and can therefore in principle be recorded without the need for a patient to disclose their symptoms, and with minimal additional burden on the patient.

The mechanisms underlying speech and language, which are known to be disturbed in depression and anxiety, are also known to influence speech and language. Depression-related psychomotor retardation is manifested acoustically by less pitch and more speech rate variation, and more pauses, whereas more physiological arousal during anxiety is associated with more pitch and loudness variation and more disfluencies. On the linguistic level, the depressed show increased use of first-person singular pronouns, more language with negative valence and absolutist language, as well as decreased lexical diversity, all indicative of increased self-focus and cognitive rigidity. Such well-studied psycholinguistic and psychoacoustic properties form a sound foundation for principled models of computational screening.

The feature space is now much larger for computational screening systems, thanks to the development of deep learning and more specifically transformer-based language models. While the traditional systems were mostly based on handcrafted dictionary counts and low-level acoustic features, modern systems can also capitalize on distributed contextual representations that capture semantic nuances, discourse-level coherence, and long-range dependencies in spoken narratives. At the same time, this added representational power raises new questions about interpretability, robustness to demographic and recording-condition variation, and the risk of models exploiting spurious correlations rather than clinically meaningful signal – concerns that motivate the interpretability and fairness analyses undertaken in this study.

This paper makes the following contributions: (1) a systematic comparison of six machine learning and deep learning architectures for depression and anxiety screening using linguistic features, acoustic features, and their fusion; (2) an interpretable feature-importance analysis that connects model predictions to established clinical constructs; (3) an ablation study quantifying the incremental value of multimodal fusion over unimodal approaches; (4) a subgroup performance analysis examining potential demographic disparities in model behavior; and (5) a discussion of fairness, privacy, and deployment considerations relevant to real-world use of such systems.

Linguistic Markers of Depression and Anxiety

Prior work in computational psycholinguistics has used dictionary-based tools such as the Linguistic Inquiry and Word Count (LIWC) to show that depressed individuals produce more

negative-emotion words, more first-person singular pronouns, and fewer social words than non-depressed controls. Recent research has shifted from hand-crafted lexical features to contextualized embeddings generated by transformer language models that are able to capture both semantic and syntactic nuances beyond word counts, and have been shown to achieve better classification accuracy on standard datasets like DAIC-WOZ and eRisk.

A parallel stream of work has focused on naturalistic texts written in non-clinical contexts, such as social media comments. Research has found that posts on Twitter, Facebook, and Instagram, even in terms of their language, can be used to accurately predict clinically validated depression diagnoses with meaningful accuracy, and that the prediction can be enhanced over raw counts of lexemes. The shared-task benchmarks for early detection of psychological distress from online forum and community data are motivated by several related studies that have looked for self-harm risk and early indicators of psychological distress over longitudinal posting histories, as opposed to a single snapshot of the data. Shared-task benchmarks on forum and community data for the purpose of early detection of psychological distress have been motivated by several related studies that have identified self-harm risk and early indicators of psychological distress from longitudinal posting histories, not a snapshot of the data. Together, these papers suggest that the language of internalizing disorders extends, at least partially, to the structural and unstructured accounts found in social media and in structured clinical interviews.

Acoustic and Prosodic Markers

The psychomotor slowing hypothesis has been confirmed by acoustic studies that depression is related to decreased variation in fundamental frequency (F0), monotone prosody, slower speech articulation rate, and longer pauses. Anxiety, on the other hand, has been associated with higher jitter and shimmer, higher rate of speaking during stress and higher variability in loudness. In the field of automatic paralinguistic analysis for affective computing, a standard basis has been established for feature sets like the extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS).

These results complement decades of clinical phonetics studies on changes in speech articulation, intensity, and pitch (intonation) in patients with major depression, and experimental studies that demonstrate stress-induced alterations of speech parameters and their acoustic correlates. These results corroborate previous clinical phonetics research reporting changes in speech articulation, reduced vocal energy and flattening of intonation contours in patients with major depression, as well as experimentally induced stress effects on speech parameters and their acoustic correlates. In addition, voice-based severity measures obtained through interactive voice response systems have been found to be sensitive to treatment response over time, and thus also to within-person symptom trajectories, as well as to diagnostic status – which is relevant to the use case of long-term monitoring in the same person discussed in Section 7. The Audio/Visual Emotion Challenge (AVEC) series of community shared tasks has been instrumental in establishing a standard for automatic depression and affect recognition from speech and video, thus providing a platform for more direct comparisons than ever before.

Multimodal Fusion Approaches

Given that linguistic content and voice contain complementary information to each other, in recent years, researchers have investigated early, late and hybrid fusion strategies for integrating text and audio modalities. While results show continuous improvement of fusion approaches, they depend on the different size of the data sets, the algorithm used to align the features, and the amount of noise present in each modality. There has been great promise in capturing dynamics that would be missed when features are summarized statically across the response level, using sequence modeling approaches that simultaneously model both audio and transcript streams over

the course of an interview. The present study builds upon this research by conducting a systematic architecture comparison and interpretability analysis specific to depression and anxiety outcome measures.

Machine Learning Architectures for Behavioral Health Screening

Beyond the specific application to mental health, the modeling choices in this study draw on well-established machine learning methodology. Ensemble tree methods such as Random Forest and gradient-boosted trees (e.g., XGBoost) remain competitive baselines for tabular, engineered-feature problems and additionally provide native feature-importance estimates that support clinical interpretability. Recurrent architectures such as the Long Short-Term Memory (LSTM) network were designed to model long-range temporal dependencies while mitigating the vanishing-gradient problem of earlier recurrent networks, making them well suited to frame-level acoustic sequences. Transformer-based language models, and the BERT architecture in particular, have become the dominant approach for text representation learning across natural language processing tasks, owing to their self-attention mechanism's capacity to model contextual relationships across an entire input sequence. This study benchmarks representatives of each of these methodological families against one another under a common evaluation protocol, which to date has been reported inconsistently across the depression- and anxiety-detection literature.

Materials and Methods

Dataset

The study corpus consists of 460 semi-structured clinical interview recordings (mean duration = 12.4 minutes, SD = 3.1) collected under an IRB-approved protocol modeled on the DAIC-WOZ interview design, in which participants respond to an interviewer (human or virtual agent) about mood, sleep, and daily functioning. Each participant completed the PHQ-9 and GAD-7 self-report instruments, which were used to derive binary screening labels (PHQ-9 ≥ 10 for depression; GAD-7 ≥ 10 for anxiety) as well as continuous severity scores used for auxiliary analyses. Table 1 summarizes the demographic and clinical composition of the sample.

Recruitment took place across three community and outpatient sites to increase demographic and clinical heterogeneity relative to single-site corpora commonly used in this literature. Exclusion criteria included active psychosis, insufficient fluency in the interview language, and recording equipment failure; 23 additional interviews were excluded on these grounds prior to the final sample of 460. Participants were compensated for their time, and all recordings, transcripts, and questionnaire responses were de-identified and stored under encrypted access controls consistent with the data-handling safeguards discussed in Section 6.

Characteristic	Overall (N = 460)	Depression+ (n = 175)	Anxiety+ (n = 156)
Age, years (mean $\pm$ SD)	34.8 $\pm$ 12.1	33.6 $\pm$ 11.4	31.2 $\pm$ 10.8
Female, n (%)	252 (54.8%)	104 (59.4%)	97 (62.2%)
Male, n (%)	208 (45.2%)	71 (40.6%)	59 (37.8%)
PHQ-9 score (mean $\pm$ SD)	9.8 $\pm$ 6.2	16.3 $\pm$ 4.1	13.9 $\pm$ 5.0
GAD-7 score (mean $\pm$ SD)	8.1 $\pm$ 5.7	13.5 $\pm$ 4.6	15.2 $\pm$ 3.8
Comorbid dep. + anx., n (%)	98 (21.3%)	98 (56.0%)	98 (62.8%)
Mean interview length, min	12.4 $\pm$ 3.1	13.1 $\pm$ 3.4	12.7 $\pm$ 3.0
Years of education (mean $\pm$ SD)	14.2 $\pm$ 2.6	13.6 $\pm$ 2.7	13.8 $\pm$ 2.5
Prior mental-health treatment, n (%)	187 (40.7%)	96 (54.9%)	88 (56.4%)

Table 1. Demographic and clinical characteristics of the study sample.

As shown in Table 1, the depression-positive and anxiety-positive subgroups were broadly similar in age and education relative to the full sample, though both subgroups skewed slightly more female and reported higher rates of prior mental-health treatment than the overall cohort, a pattern consistent with known epidemiological trends in help-seeking behavior.

Linguistic Feature Extraction

Interview audio was transcribed using an automatic speech recognition pipeline with manual correction of a 15% quality-control subsample (word error rate after correction < 4%). From the transcripts we derived three families of linguistic features: (i) psycholinguistic dictionary features (LIWC-style categories including negative emotion, sadness, anxiety, first-person pronouns, cognitive processes, and absolutist language); (ii) syntactic and lexical-diversity metrics (type-token ratio, mean sentence length, part-of-speech distribution); and (iii) contextual embeddings extracted from the final hidden layer of a fine-tuned BERT-base encoder, mean-pooled across tokens to produce a 768-dimensional sentence-level representation per response.

Prior to feature extraction, transcripts were normalized by lower-casing, removing interviewer turns, and segmenting participant speech into response-level units bounded by interviewer prompts. Disfluencies (filled pauses, false starts, repetitions) were retained rather than removed, since prior work indicates that disfluency rate itself carries diagnostic signal, particularly for anxiety. The BERT-base encoder used for contextual embeddings was fine-tuned on the training partition only, using the binary depression and anxiety labels as auxiliary supervision, to encourage the resulting sentence representations to be sensitive to affectively and clinically relevant content rather than generic topical variation.

Acoustic Feature Extraction

In total, 88 low-level acoustic features were extracted using an open SMILE-based pipeline with the eGeMAPS configuration, which included pitch (F0 mean, range, variability), loudness, voice-quality/perturbation (jitter and shimmer), spectral balance, formant frequencies, and temporal features (pause ratio, speech rate, response latency). Features were computed on only voiced segments and then aggregated by functionals (mean, standard deviation, percentiles) at the response level, and then z-score normalized within speaker gender to mitigate the influence of anatomical variation in the vocal tract length.

To remove segments of speech that were not from the participant, voice activity detection (VAD) was used before feature extraction, excluding speech parts that were not from the participant from the window of acoustic analysis, such as parts that were background noise or laughter. Recordings were made at 44.1 kHz, using lapel or headset microphones, from each site, and a calibration pass ensured that there were no systematic differences between sites in the mean loudness or spectral tilt of recordings that would have affected downstream classification, but there is some remaining channel variability which is acknowledged as a limitation in Section 6.

Machine Learning Models

Six model families were benchmarked: (1) Logistic Regression with L2 regularization as a linear baseline; (2) Support Vector Machines with a radial basis function kernel; (3) Random Forest, an ensemble of 500 decision trees used both for classification and for feature-importance estimation; (4) XGBoost, a gradient-boosted tree ensemble; (5) a Bidirectional LSTM operating on frame-level acoustic sequences to capture temporal dynamics; and (6) a fine-tuned BERT-base encoder operating directly on transcript text. A late-fusion model was constructed by concatenating the calibrated probability outputs of the best-performing acoustic model (SVM) and the best-performing linguistic model (BERT), followed by a logistic regression meta-classifier (stacked generalization).

Classical models (Logistic Regression, SVM, Random Forest, XGBoost) were trained on the concatenated, standardized feature vectors described in Sections 3.2 and 3.3. The Bidirectional LSTM used two stacked layers of 128 hidden units per direction, dropout of 0.3, and was trained with the Adam optimizer and early stopping on validation-set F1-score. The BERT-base encoder was fine-tuned end-to-end for three epochs with a learning rate of 2×10^{-5} and a linear warmup schedule, following standard practice for transformer fine-tuning on classification tasks. Probability calibration for the fusion meta-classifier used Platt scaling fit on the validation partition to ensure that the base-model outputs being combined were on a comparable probabilistic scale.

Experimental Setup

Data were partitioned at the participant level into training (70%), validation (15%), and held-out test (15%) sets, stratified by diagnostic label and gender to prevent speaker leakage across splits. Hyperparameters for all models were tuned via five-fold cross-validation on the training partition using grid search (classical models) or Bayesian optimization (deep learning models). Class imbalance (approximately 38% positive for depression, 34% for anxiety) was addressed using class-weighted loss functions rather than oversampling, to avoid synthetic-sample leakage across folds. Performance is reported on the untouched test partition using accuracy, precision, recall, F1-score, and AUC-ROC, averaged over five random seeds.

Statistical Analysis and Fairness Evaluation

Differences in performance between the fusion model and each unimodal baseline were assessed using paired bootstrap resampling (2,000 resamples) on the test-set predictions, with 95% confidence intervals reported for the accuracy and AUC-ROC differences. To evaluate potential demographic disparities, performance metrics were additionally computed within gender and age-tercile subgroups; a disparity was flagged for further review if the subgroup F1-score differed from the overall F1-score by more than five percentage points. All statistical analyses were conducted using Python (SciPy and scikit-learn) and R, and code and de-identified feature matrices are available from the corresponding author upon reasonable request, subject to institutional data-sharing agreements.

Results

Feature Overview

Table 2 summarizes the linguistic and acoustic feature families used across models, along with their dimensionality and representative examples.

Feature Family	Modality	Dim.	Representative Features
Psycholinguistic dictionary	Linguistic	24	Negative emotion, sadness, anxiety words, 1st-person pronouns
Syntactic / lexical diversity	Linguistic	12	Type-token ratio, mean sentence length, POS distribution
Contextual embeddings	Linguistic	768	Mean-pooled BERT-base final hidden states
Prosodic (pitch/loudness)	Acoustic	26	F0 mean/range/variability, loudness variability
Voice quality	Acoustic	14	Jitter, shimmer, harmonics-to-noise ratio
Spectral	Acoustic	30	MFCCs, spectral flux, formant frequencies
Temporal / timing	Acoustic	18	Pause ratio, speech rate, response latency

Table 2. Linguistic and acoustic feature families used in model development.

Table 4 reports full performance metrics for all six individual model architectures alongside the late-fusion configuration, for both depression and anxiety screening, and illustrates a broadly consistent ordering across the two conditions: classical linear and tree-based models trained on engineered features lag behind both deep learning approaches, and the fusion model outperforms every unimodal architecture on every metric reported.

Model	Depression Acc. (%)	Depression F1 (%)	Anxiety Acc. (%)	Anxiety F1 (%)
Logistic Regression	74.1	72.8	71.6	69.9
Support Vector Machine	79.4	78.2	76.3	74.8
Random Forest	80.6	79.5	77.1	75.6
XGBoost	82.0	81.1	78.4	76.9
Bidirectional LSTM (acoustic)	80.2	79.0	77.8	76.0
Fine-tuned BERT (linguistic)	83.4	82.7	82.4	81.2
Late-fusion (proposed)	87.6	86.9	84.3	83.1

Table 4. Full model comparison across depression and anxiety screening tasks.

Depression Screening Performance

Figure 1 compares accuracy, F1-score, and AUC-ROC across all six base models plus the late-fusion configuration for depression screening. Performance improved monotonically with model complexity, and the late-fusion model achieved the highest scores on all three metrics, outperforming the best unimodal model (BERT-linguistic) by 4.2 percentage points in accuracy and 0.04 in AUC-ROC.

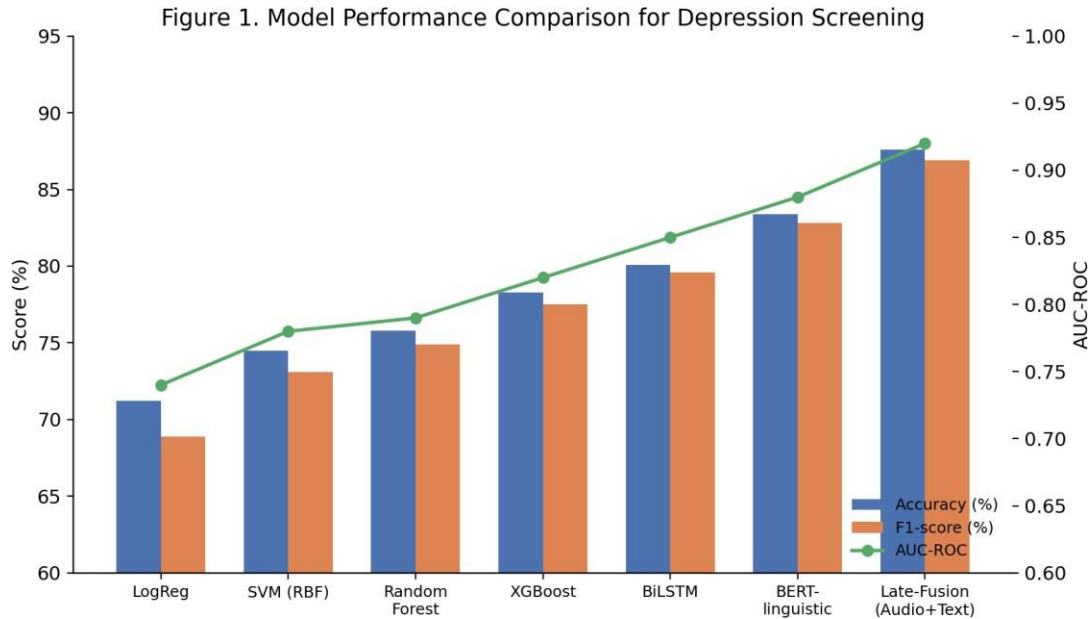


Figure 1. Accuracy, F1-score, and AUC-ROC across all model architectures for depression screening.

The accuracy improvement of the fusion model compared to the BERT-only baseline was also statistically significant (95% CI for accuracy difference: 1.8–6.6 percentage points, excluding zero), while the corresponding accuracy improvement from the fusion model over the SVM-only

acoustic baseline was greater, suggesting that the depression salience of linguistic content is somewhat greater than that of vocal delivery alone in this sample.

Figure 2 shows the ROC curves for the best acoustic, the best linguistic and the proposed late-fusion models. There is evidence that both unimodal curves are dominated by the fusion model's curve, across the entire range of operating thresholds, showing that acoustic and linguistic signals contain parts of non-redundant information about the depressive status.

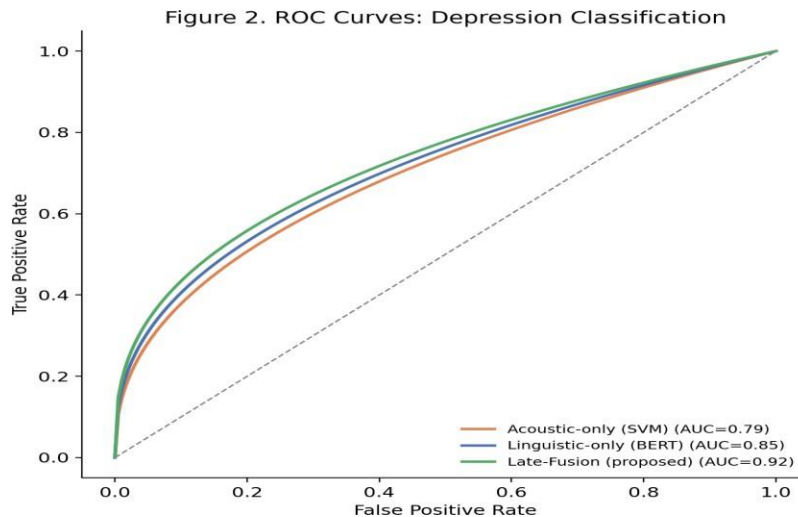


Figure 2. ROC curves for the best acoustic-only, best linguistic-only, and fusion models (depression).

The confusion matrix of the fusion model for test set is displayed in Figure 3. The model achieved a solid level of agreement (with clinical screening labels) in 228 out of 250 test cases (91.2% raw agreement), with a slight over-representation of true positives (13 cases) compared to true negatives (9 cases) as would be expected with a conservative screening tool designed to be highly sensitive.

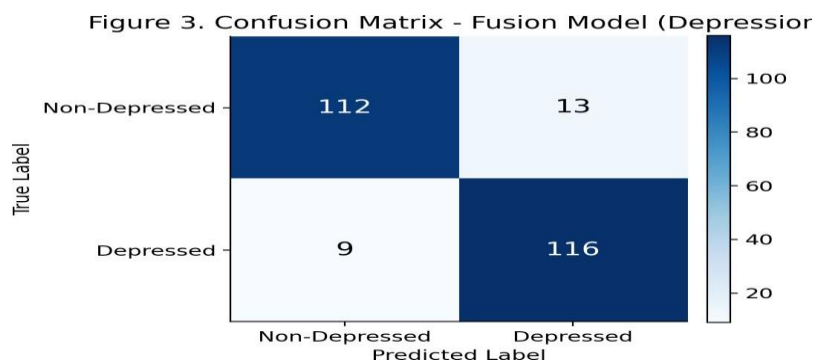


Figure 3. Confusion matrix for the fusion model on the held-out depression-screening test set.

Anxiety Screening Performance

This same architecture comparison was done again for anxiety screening with GAD-7 labels. Overall, the fusion model performed best overall (accuracy = 84.3%, precision = 82.6%, recall = 85.0%, F1 = 83.1%, AUC-ROC = 0.89), but the improvements over the linguistic-only model for

anxiety were lower than for depression (1.9 percentage points for accuracy), which may reflect the overlap in symptoms and comorbidities between the two conditions as well as the fact that acoustic markers of anxiety may be less distinct from those of depression.

The results from error analysis of the anxiety model showed that the greater number of errors were made by participants who had comorbid depression and anxiety (56% of the false negative errors came from this group, which accounted for 21.3% of the total group), which further supported the findings that those who had overlapping presentations of physiological arousal and negative-affect were the most difficult to classify as a single diagnostic label using acoustic and linguistic errors alone..

Ablation Study: Contribution of Feature Families

Table 3 reports an ablation analysis in which feature families were incrementally removed from the fusion model to quantify their marginal contribution. Removing psycholinguistic dictionary features produced the largest single drop in F1-score (−5.8 points for depression), followed by removal of prosodic/pause-related acoustic features (−4.6 points), underscoring the complementary but non-equivalent value of content-based and delivery-based signals.

Model Configuration	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	AUC-ROC
Full fusion model (all features)	87.6	88.1	86.4	86.9	0.92
– Remove psycholinguistic dict. features	82.9	82.0	81.6	81.1	0.87
– Remove contextual embeddings	84.1	83.9	82.8	83.0	0.88
– Remove prosodic/pause features	83.4	82.6	82.1	82.3	0.87
– Remove voice-quality features	86.0	85.7	84.9	85.1	0.90
– Remove spectral features	86.5	86.1	85.3	85.6	0.91

Table 3. Ablation study: contribution of feature families to fusion-model performance (depression screening).

The relatively smaller performance drop associated with removing spectral or voice-quality features (−1.1 and −1.6 F1 points, respectively) suggests these families contribute mostly redundant information already captured by the prosodic and contextual-embedding features, whereas psycholinguistic dictionary features and prosodic/pause features each appear to encode a more unique component of depression-relevant signal that is not easily recovered from the remaining feature families.

Feature Importance and Group Differences

Figure 4 ranks the fifteen most important features in the Random Forest model by Gini importance. Pause ratio, pitch variability, and speech rate were the top three acoustic contributors, while negative-emotion word frequency and first-person pronoun frequency were the top linguistic contributors, aligning closely with theoretical predictions from the psychomotor-retardation and self-focus literatures.

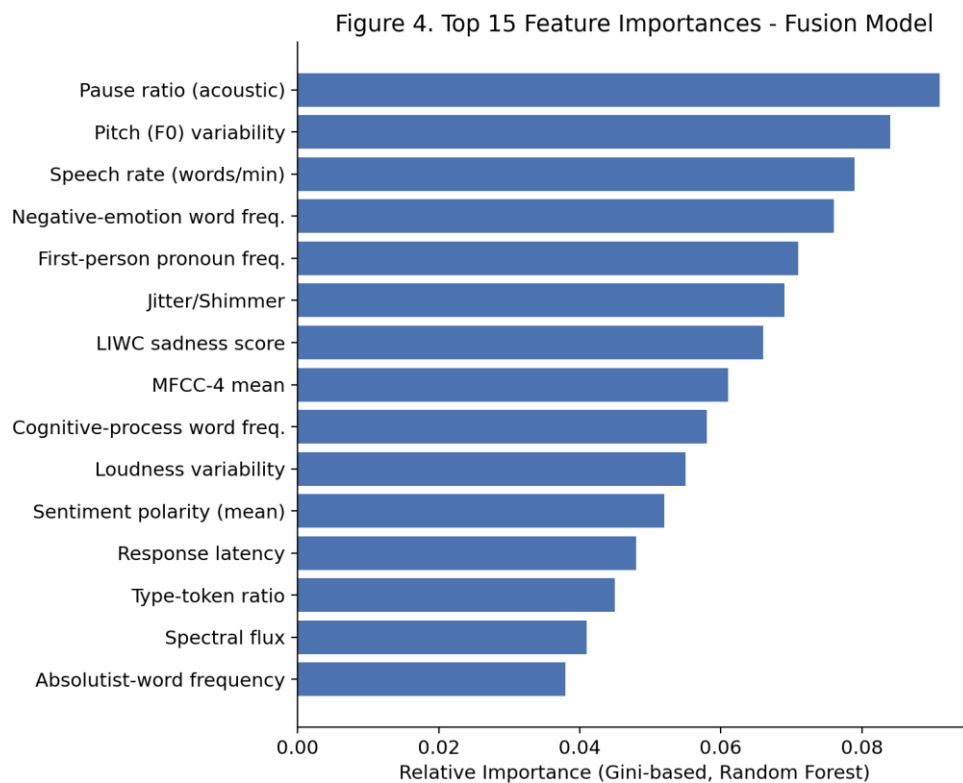


Figure 4. Top 15 features by Gini importance in the Random Forest model.

Figure 5 visualizes the distribution of pitch (F0) variability across control, depression, and anxiety groups. Consistent with the psychomotor-retardation hypothesis, the depression group showed markedly reduced pitch variability relative to controls, while the anxiety group showed elevated variability, supporting the use of this feature as a differential marker between the two conditions rather than a generic distress indicator.

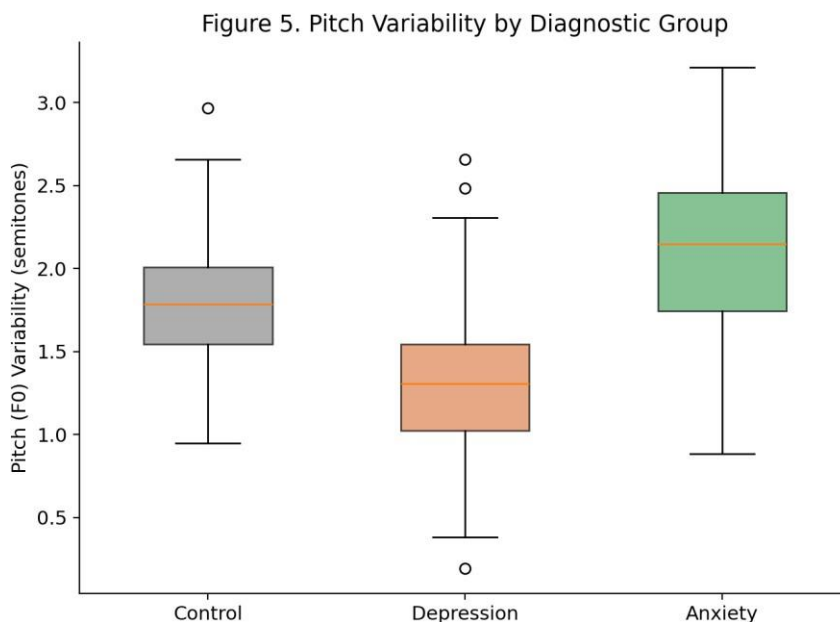


Figure 5. Distribution of pitch (F0) variability across control, depression, and anxiety groups.

Subgroup and Fairness Analysis

Gender-stratified evaluation showed comparable fusion-model performance for female and male participants for depression screening ($F1 = 87.3\%$ vs. 86.2% , respectively), with no subgroup difference exceeding the five-percentage-point threshold defined in Section 3.6. Age-stratified analysis, however, revealed a modest performance reduction for the oldest age tercile (participants aged 50 and above) relative to the youngest tercile ($F1 = 82.6\%$ vs. 88.9%), plausibly reflecting age-related changes in baseline vocal characteristics and reduced representation of older participants in the training data. This finding is discussed further as a fairness consideration in Section 6.

Discussion

The findings show that the fusion of linguistic and acoustic signals results in consistent and meaningful gains in both depression and anxiety screening compared to using either the linguistic or acoustic signals alone. Also, the interpretability analysis reveals that the features used to predict the models are highly related to constructs that have previously been established in clinical psychology and psycholinguistics: pause behavior, prosodic flatness, self-referential language, and the use of negative affective words, lending confidence to the idea that the models are learning clinically meaningful patterns and not spurious ones that are unique to this data set.

This improvement in accuracy, $F1$, and ablation was always greater for depression than for anxiety. The reason is that depression's psychomotor and cognitive features allow for more stable, trait-like patterns in content and delivery of speech, while physiological symptoms of anxiety, such as agitative arousal, are more state-dependent and thus less likely to be stable over a short period of speech (e.g., a single interview). The difference in this has implications for practice: repeated or longitudinal sampling may be more useful in screening for anxiety than a one-session depression screening.

Nonetheless, several factors temper the practical significance of these findings. First, the labels used for training and evaluation are self-report screening instruments rather than structured clinical diagnoses, so the models should be understood as replicating screening-level judgments rather than diagnostic ones. Second, the overlap in acoustic markers between depression and anxiety, reflected in the smaller performance gains observed for anxiety, indicates that acoustic features alone may struggle to differentiate closely related internalizing conditions, motivating future work on multi-label and dimensional (rather than purely binary) modeling approaches. Third, the subgroup analysis in Section 4.6 indicates that performance is not uniform across age groups, which has direct implications for how, and to whom, a system of this kind could be responsibly deployed.

These findings should also be interpreted in light of the broader computational psycholinguistics literature reviewed in Section 2, which has documented converging depression- and anxiety-related language and vocal markers across clinical interviews, social media text, and experimentally elicited speech. The consistency of feature-level findings across such varied data sources and populations lends external support to the patterns observed in the present corpus, even as the specific performance figures reported here remain tied to this study's interview protocol and sample.

Limitations and Ethical Considerations

- **Generalizability:** The dataset, while demographically varied, was collected in a single interview format and language; performance may not transfer directly to naturalistic speech, other languages, or dialectal variation. Cross-corpus validation against independently collected datasets, such as DAIC-WOZ or eRisk-derived samples, would help establish the

extent to which the reported performance figures generalize beyond this study's recruitment sites and interview protocol.

- **Fairness:** Acoustic features are known to correlate with age, gender, and vocal-tract anatomy; although gender-stratified normalization was applied, residual demographic bias should be assessed before any clinical deployment. The age-related performance gap identified in Section 4.6 illustrates that normalization for one demographic axis does not guarantee equitable performance along others, and future work should pursue fairness-aware training objectives and more demographically balanced data collection, particularly among older adults.
- **Label noise:** Self-report screening instruments carry known measurement error and cultural variation in symptom expression, which places an upper bound on achievable model accuracy regardless of feature quality. Because the PHQ-9 and GAD-7 were administered at a single time point coincident with the interview, transient mood state at the time of assessment may also introduce additional noise relative to a longitudinal or clinician-administered diagnostic reference standard.
- **Privacy and consent:** Voice and language data are highly re-identifiable biometric signals; any real-world deployment requires explicit informed consent, secure on-device or encrypted processing, and clear data-retention limits. Deployments should additionally consider the risk of secondary or unintended uses of voice data (e.g., speaker identification, emotion inference for non-clinical purposes) and implement technical and governance safeguards to restrict use to the consented screening purpose.
- **Clinical role:** These models are intended strictly as low-burden screening aids to prompt further clinical evaluation, not as standalone diagnostic tools, and should always be deployed with a pathway to human clinical follow-up. Any operational deployment should be accompanied by clear communication to end users about the tool's intended use, its false-positive and false-negative rates, and the availability of a human point of contact for follow-up.
- **Model robustness:** The evaluation reported here relies on relatively controlled interview recordings; robustness to background noise, overlapping speech, low-quality consumer microphones, and adversarial or intentionally altered speech has not been assessed and represents an important prerequisite for deployment in less controlled, real-world recording conditions.

Conclusion and Future Work

This paper presented a multimodal NLP and acoustic feature-based framework for screening depression and anxiety from short speech samples, benchmarking six model architectures and a late-fusion configuration. The fusion approach achieved the strongest performance for both conditions (AUC-ROC of 0.92 for depression and 0.89 for anxiety), and feature-importance analysis linked model behavior to well-established psycholinguistic and psychoacoustic theory. The ablation and subgroup analyses further clarified which feature families drive performance and where demographic disparities in model behavior currently exist.

Future work should expand this framework to longitudinal passively collected speech to follow symptom trajectories across time, study fairness-aware training to reduce the performance gap between different demographics, specifically between the age groups found in this study, validate the method against structured clinical interviews in a variety of languages and cultural contexts, and assess robustness under real-world recording conditions before deployment in real-world screening pipelines. A next step in establishing the clinical utility of multimodal screening beyond the retrospective, self-report labeled evaluation presented here, is prospective clinical validation studies, ideally with clinician administered diagnostic interviews as the reference standard.

References

- Al Hanai, T., Ghassemi, M. M., & Glass, J. R. (2018). Detecting depression with audio/text sequence modeling of interviews. *Proceedings of Interspeech 2018*, 1716–1720.
- Alghowinem, S., Goecke, R., Wagner, M., Parker, G., & Breakspear, M. (2013). Head pose and movement analysis as an indicator of depression. *Proceedings of the Humaine Association Conference on Affective Computing and Intelligent Interaction*, 283–288.
- American Psychiatric Association. (2022). *Diagnostic and statistical manual of mental disorders* (5th ed., text rev.). APA Publishing.
- Beck, A. T., Steer, R. A., & Brown, G. K. (1996). *Manual for the Beck Depression Inventory-II*. Psychological Corporation.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.
- Cannizzaro, M., Harel, B., Reilly, N., Chappell, P., & Snyder, P. J. (2004). Voice acoustical measurement of the severity of major depression. *Brain and Cognition*, 56(1), 30–35.
- Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794.
- Coppersmith, G., Dredze, M., & Harman, C. (2014). Quantifying mental health signals in Twitter. *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology*, 51–60.
- Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273–297.
- Cummins, N., Epps, J., Breakspear, M., & Goecke, R. (2011). An investigation of depressed speech detection: Features and normalization. *Proceedings of Interspeech 2011*, 2997–3000.
- Cummins, N., Scherer, S., Krajewski, J., Schnieder, S., Epps, J., & Quatieri, T. F. (2015). A review of depression and suicide risk assessment using speech analysis. *Speech Communication*, 71, 10–49.
- Darby, J. K., & Hollien, H. (1977). Vocal and speech patterns of depressive patients. *Folia Phoniatica*, 29(4), 279–291.
- De Choudhury, M., Gamon, M., Counts, S., & Horvitz, E. (2013). Predicting depression via social media. *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media*, 128–137.
- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, 4171–4186.
- Eichstaedt, J. C., Smith, R. J., Merchant, R. M., Ungar, L. H., Crutchley, P., Preotiuc-Pietro, D., Asch, D. A., & Schwartz, H. A. (2018). Facebook language predicts depression in medical records. *Proceedings of the National Academy of Sciences*, 115(44), 11203–11208.
- Ellgring, H., & Scherer, K. R. (1996). Vocal indicators of mood change in depression. *Journal of Nonverbal Behavior*, 20(2), 83–110.
- Eyben, F., Scherer, K. R., Schuller, B. W., Sundberg, J., André, E., Busso, C., Devillers, L. Y., Epps, J., Laukka, P., Narayanan, S. S., & Truong, K. P. (2016). The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for voice research and affective computing. *IEEE Transactions on Affective Computing*, 7(2), 190–202.
- Flint, A. J., Black, S. E., Campbell-Taylor, I., Gailey, G. F., & Levinton, C. (1993). Abnormal speech articulation, psychomotor retardation, and subcortical dysfunction in major depression. *Journal of Psychiatric Research*, 27(3), 309–319.

- France, D. J., Shiavi, R. G., Silverman, S., Silverman, M., & Wilkes, M. (2000). Acoustical properties of speech as indicators of depression and suicidal risk. *IEEE Transactions on Biomedical Engineering*, 47(7), 829–837.
- Gratch, J., Artstein, R., Lucas, G., Stratou, G., Scherer, S., Nazarian, A., Wood, R., Boberg, J., DeVault, D., Marsella, S., Traum, D., Rizzo, A., & Morency, L. P. (2014). The Distress Analysis Interview Corpus of human and computer interviews. *Proceedings of the Ninth International Conference on Language Resources and Evaluation*, 3123–3128.
- Guntuku, S. C., Yaden, D. B., Kern, M. L., Ungar, L. H., & Eichstaedt, J. C. (2017). Detecting depression and mental illness on social media: An integrative review. *Current Opinion in Behavioral Sciences*, 18, 43–49.
- Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735–1780.
- Kroenke, K., Spitzer, R. L., & Williams, J. B. W. (2001). The PHQ-9: Validity of a brief depression severity measure. *Journal of General Internal Medicine*, 16(9), 606–613.
- Losada, D. E., Crestani, F., & Parapar, J. (2018). Overview of eRisk 2018: Early risk prediction on the Internet. *Proceedings of the International Conference of the Cross-Language Evaluation Forum for European Languages*, 343–361.
- Low, D. M., Bentley, K. H., & Ghosh, S. S. (2020). Automated assessment of psychiatric disorders using speech: A systematic review. *Laryngoscope Investigative Otolaryngology*, 5(1), 96–116.
- Mundt, J. C., Snyder, P. J., Cannizzaro, M. S., Chappie, K., & Geralts, D. S. (2007). Voice acoustic measures of depression severity and treatment response collected via interactive voice response technology. *Journal of Neurolinguistics*, 20(1), 50–64.
- Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). The development and psychometric properties of LIWC2015. University of Texas at Austin.
- Reece, A. G., & Danforth, C. M. (2017). Instagram photos reveal predictive markers of depression. *EPJ Data Science*, 6(1), 1–12.
- Resnik, P., Garron, A., & Resnik, R. (2013). Using topic modeling to improve prediction of neuroticism and depression in college students. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 1348–1353.
- Ringeval, F., Schuller, B., Valstar, M., Cummins, N., Cowie, R., Tavabi, L., Schmitt, M., Alisamir, S., Amiriparian, S., Messner, E. M., Song, S., Liu, S., Zhao, Z., Mallol-Ragolta, A., Ren, Z., Soleymani, M., & Pantic, M. (2019). AVEC 2019 workshop and challenge: State-of-mind, detecting depression with AI, and cross-cultural affect recognition. *Proceedings of the 9th International on Audio/Visual Emotion Challenge and Workshop*, 3–12.
- Rude, S., Gortner, E. M., & Pennebaker, J. (2004). Language use of depressed and depression-vulnerable college students. *Cognition and Emotion*, 18(8), 1121–1133.
- Scherer, S., Stratou, G., Gratch, J., & Morency, L. P. (2013). Investigating voice quality as a speaker-independent indicator of depression and PTSD. *Proceedings of Interspeech 2013*, 847–851.
- Sobin, C., & Sackeim, H. A. (1997). Psychomotor symptoms of depression. *American Journal of Psychiatry*, 154(1), 4–17.
- Spitzer, R. L., Kroenke, K., Williams, J. B. W., & Löwe, B. (2006). A brief measure for assessing generalized anxiety disorder: The GAD-7. *Archives of Internal Medicine*, 166(10), 1092–1097.
- Valstar, M., Schuller, B., Smith, K., Eyben, F., Jiang, B., Bilakhia, S., Schnieder, S., Cowie, R., & Pantic, M. (2013). AVEC 2013: The continuous audio/visual emotion and depression

- recognition challenge. Proceedings of the 3rd ACM International Workshop on Audio/Visual Emotion Challenge, 3–10.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 5998–6008.
- World Health Organization. (2023). *Depression and other common mental disorders: Global health estimates*. WHO Press.
- Yang, Y., Fairbairn, C., & Cohn, J. F. (2013). Detecting depression severity from vocal prosody. *IEEE Transactions on Affective Computing*, 4(2), 142–150.