Automated Mood Detection from Naturalistic Phone Conversations: A Feasibility Study for Relapse Prediction in Mood Disorders

Rajib Rana, Niall Seamus Higgins, Kazi Nazmul Haque, Björn Wolfgang Schuller

Submitted to: JMIR Mental Health
on: December 16, 2025

Table of Contents

Automated Mood Detection from Naturalistic Phone Conversations: A Feasibility Study for Relapse Prediction in Mood Disorders

Rajib Rana[1*] PhD; Niall Seamus Higgins[2*] PhD; Kazi Nazmul Haque[1*] PhD; Björn Wolfgang Schuller[3*] PhB

[1] School of Mathematics, Physics and Computing University of Southern Queensland Springfield Central AU
[2] The Park - Mental Health and Specialised Services West Moreton Health Wacol AU
[3] Technical University of Munich Imperial College London London GB
[*]these authors contributed equally

Corresponding Author:
Rajib Rana PhD

School of Mathematics, Physics and Computing
University of Southern Queensland
37 Sinnathamby Boulevard
Springfield Central
AU

Abstract

Background: Mood disorders have high relapse rates and existing monitoring relies on infrequent clinical assessments and self-report, limiting timely intervention

Objective: To evaluate the feasibility of a smartphone-based system that passively infers mood from naturalistic phone conversations using speech signal processing and artificial intelligence, and to examine alignment with self-reported mood and clinical measures

Methods: We deployed a background smartphone app to capture speech during routine calls and prompted post-call mood ratings. Encrypted features (speaker diarisation, prosody, Mel-Frequency Cepstral Coefficients (MFCCs), Word2Vec embeddings) were processed on secure servers. Phase 1 validated inferred mood against post-call self-ratings; Phase 2 compared daily mood trajectories with the Montgomery–Åsberg Depression Rating Scale (MADRS) and Early Warning Signs Questionnaire (EWSQ). The MADRS, routinely used in the hospital service, was employed to maintain continuity with clinical practice and minimise disruption to workflows.

Results: Eleven participants completed Phase 1. The inferred mood demonstrated a moderate correlation with self-reported ratings, with performance improving as call volumes increased. The pipeline operated across heterogeneous devices and preserved privacy via feature-vector transmission

Conclusions: Speech-based mood inference from naturalistic phone calls is feasible and aligns with subjective and clinical indicators, especially with sufficient call activity. Privacy-preserving design and multimodal features facilitate real-world deployment while promoting proactive relapse prevention. Clinical Trial: No Trial

(JMIR Preprints 16/12/2025:89660)
DOI: https://doi.org/10.2196/preprints.89660

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

✓ **Please make my preprint PDF available to anyone at any time (recommended).**
Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v
Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http
No. Please do not make my accepted manuscript PDF available to anyone. I understand that if I later pay to participate in <a href="https:/

Original Manuscript

Original paper

Automated Mood Detection from Naturalistic Phone Conversations: A Feasibility Study for Relapse Prediction in Mood Disorders

Rajib Rana, PhD[1]; Niall Higgins, PhD[1,2]; Kazi N. Haque, PhD[1]; and Björn W. Schuller, PhD[3,4]

[1]School of Mathematics, Physics and Computing, University of Southern Queensland Education City, Springfield Central, Australia
[2]The Park - Mental Health and Specialised Services, West Moreton Health, Wacol, QLD, Australia
[3]GLAM – the Group on Language, Audio, & Music, Imperial College London, UK
[4]Head of CHI - Chair of Health Informatics, Technical University of Munich (TUM), Munich, Germany

Corresponding Author:
Rajib Rana
School of Mathematics, Physics and Computing
Office B341, Springfield Campus
University of Southern Queensland
Springfield, QLD, Australia
Email: Rajib.Rana@unisq.edu.au

Abstract

Background: Mood disorders have high relapse rates and existing monitoring relies on infrequent clinical assessments and self-report, limiting timely intervention.

Objective: To evaluate the feasibility of a smartphone-based system that passively infers mood from naturalistic phone conversations using speech signal processing and artificial intelligence, and to examine alignment with self-reported mood and clinical measures.

Methods: We deployed a background smartphone app to capture speech during routine calls and prompted post-call mood ratings. Encrypted features (speaker diarisation, prosody, Mel-Frequency Cepstral Coefficients (MFCCs), Word2Vec embeddings) were processed on secure servers. Phase 1 validated inferred mood against post-call self-ratings; Phase 2 compared daily mood trajectories with the Montgomery–Åsberg Depression Rating Scale (MADRS) and Early Warning Signs Questionnaire (EWSQ). The MADRS, routinely used in the hospital service, was employed to maintain continuity with clinical practice and minimise disruption to workflows.

Results: Eleven participants completed Phase 1. The inferred mood demonstrated a moderate correlation with self-reported ratings, with performance improving as call volumes increased. The pipeline operated across heterogeneous devices and preserved privacy via feature-vector transmission.

Conclusions: Speech-based mood inference from naturalistic phone calls is feasible and aligns with subjective and clinical indicators, especially with sufficient call activity. Privacy-preserving design and multimodal features facilitate real-world deployment while promoting proactive relapse prevention.

KEYWORDS: Mood disorders; relapse prediction; smartphone sensing; speech analysis; artificial intelligence; mental health.

Introduction

Background

Mood disorders, including major depressive disorder and bipolar disorder, are among the leading causes of disability worldwide and represent a significant public health challenge [1]. The Global Burden of Disease Study reported that mental and substance use disorders accounted for a substantial proportion of years lived with disability, with mood disorders contributing the largest share [1]. Despite advances in pharmacological and psychological treatments, relapse remains common, with estimates suggesting that up to 80% of individuals who have experienced two episodes of depression will relapse again [2].

Major depressive disorder affects approximately one in six men and one in four women during their lifetime [3], and bipolar disorder, though less prevalent, is associated with severe functional impairment and high recurrence rates [4]. The recurrent nature of these conditions imposes a heavy burden on individuals, families, and health systems, including increased risk of suicide [5] and substantial economic costs due to lost productivity and healthcare utilisation [6].

Early detection of relapse is critical because symptom deterioration often occurs gradually over several weeks, providing a window for timely intervention [7]. However, current relapse prevention strategies rely heavily on self-report and scheduled clinical assessments, which are limited by poor insight, stigma, and restricted access to services [8]. Furthermore, traditional monitoring approaches fail to capture real-time changes in mood, leaving clinicians without objective data between appointments.

Recent advances in mobile health and artificial intelligence (AI) offer promising solutions for continuous, passive monitoring of mental health status. Speech is a rich source of emotional and cognitive information; changes in prosody, pitch, and linguistic patterns often precede clinical deterioration [9, 10]. By leveraging these features through machine learning, automated mood inference from naturalistic phone conversations becomes possible, offering an objective and scalable approach to relapse prediction.

This study evaluated the feasibility of a privacy preserving, smartphone-based system that infers mood from naturalistic phone conversations using AI-driven speech analysis. The system processes only encrypted, de-identified acoustic features, never raw audio or linguistic content, ensuring participant confidentiality while enabling automated mood inference. Analyses focused on whether the system could generate consistent mood estimates from de-identified numerical outputs, aligning these with self-reported and clinical measures, all within a secure research environment that safeguards participant privacy. A fine-tuned wav2vec model generated emotion probability sequences from speech, which were then mapped to mood categories using a proprietary algorithm. A detailed account of the analytical pipeline is included in the Methods section.

Methods

Study Design

This feasibility study was structured in two phases:

- **Phase 1:** Validate the mood detection algorithm against participants' self-reported mood ratings during routine phone calls.
- **Phase 2:** Map inferred mood states to clinical assessment scores, including the MADRS and EWSQ, in participants with a confirmed diagnosis of a mood disorder.

We deployed a background smartphone app to capture speech during routine calls and prompted post-call mood ratings. Encrypted features, including speaker diarisation, prosody,

MFCCs, and Word2Vec embeddings, were processed on secure servers. MFCCs, widely used in both clinical and research settings, effectively capture vocal timbre and emotional expression [11]. Phase 1 validated inferred mood against post-call self-ratings, while Phase 2 compared daily mood trajectories with the MADRS and EWSQ. The MADRS, a clinician rated instrument consisting of ten items scored from 0 to 6 (total range 0–60), Is responsive to short-term changes in symptoms and is frequently used to measure treatment response. Unlike patient-reported measures such as the PHQ-9 or BDI, it exhibits lower vulnerability to ceiling effects, maintaining utility across severity levels and minimising extra assessments to support efficient clinical processes. As the standard assessment tool within the hospital service, MADRS scores were collected at baseline during the consent visit and, when available, at follow-up to contextualise inferred mood trajectories and validate the mood inference algorithm.

Participants

Participants were recruited from Metro North Mental Health outpatient services in Brisbane, Queensland, Australia. This setting was chosen because it provides ongoing support to individuals with mood-related symptoms, ensuring access to a clinically relevant population while maintaining treatment stability.

Inclusion Criteria: Eligible participants were required to have a documented history of mood symptoms (e.g., depressive or bipolar features), demonstrate stability in their current treatment regimen (defined by clinical judgement of the treating team), and engage in frequent phone communication (Average ≥ 1 outgoing or incoming call per day over the past 2 weeks), as the system relies on naturalistic voice data. Participants also needed to own a compatible smartphone capable of running the study application.

Exclusion Criteria: Individuals were excluded if they exhibited active substance use within the previous 10 days, had a diagnosis of psychosis, severe anxiety disorder, or personality disorder, or were unable to provide informed consent. Additional exclusions included insufficient phone usage (fewer than 2–3 calls per day) and lack of suitable hardware.

Recruitment was conducted by mental health Registered Nurses, clinical psychologists, and social workers during routine appointments to assess suitability for the study. This process ensured participants were clinically stable and able to engage safely with the protocol. All participants provided informed consent and were advised that they could withdraw at any time without any impact on their clinical care.

Data Analysis

All analyses were conducted within the Queensland Cyber Infrastructure Foundation (QCIF) secure research environment using only encrypted, de-identified numerical data. The primary aim was to assess the consistency and validity of system generated mood estimates derived from uplink only emotion probability sequences, produced by a fine-tuned wav2vec model, and the corresponding call level mood classifications from a proprietary heuristic algorithm.

Textbox 1: System Components

Smartphone application

- Runs continuously in the background on the participant’s device.
- Captures voice samples during routine phone calls without disrupting normal usage.
- Prompts participants to rate their mood immediately after the call using icons (Positive, Neutral, Negative).
- Collects initial voice samples during setup by asking participants to read a short text for 2–3 minutes, enabling accurate speaker diarisation.

Server component

- Processes encrypted speech features using AI algorithms on secure QCIF infrastructure.
- Converts audio into encrypted feature vectors before transmission to ensure privacy.
- Performs speaker diarisation, feature extraction (Filter Bank, MFCC, Word2Vec), and emotion classification.
- Generates call-level mood scores and aggregates them into daily mood trajectories.

For

each call, the wav2vec model produced a probability distribution across Ekman's six basic emotions. To derive a single mood label for each call, these emotion probabilities were mapped onto a valence axis according to their affective polarity (for example, joy was mapped as positive, while sadness and anger were mapped as negative). A weighted average of these valence mapped probabilities was then calculated across the duration of the call, resulting in a single call-level mood label (positive, neutral, or negative). This approach enabled the transformation of complex, temporal emotion profiles into a concise summary of overall mood for each call. Internal stability of mood estimates was assessed across individual calls, and alignment with participants' self-reported post-call ratings was evaluated using correlation analyses based solely on de-identified values.

To examine mood dynamics over time, call level mood estimates were aggregated into daily trajectories using a weighted function accounting for call frequency and duration. These daily profiles were compared with self-reported mood and, where available, with clinical assessments (MADRS and EWSQ). The analysis emphasised convergence of temporal trends rather than point-to-point prediction, reflecting the feasibility focus of the study.

Feasibility was further assessed by examining participant engagement, device differences, and system performance in everyday use. Metrics included the number of calls per participant, the distribution of call durations, and compliance with post call self-ratings. All procedures adhered to ethics protocols, ensuring no possibility of reconstructing speech content.

Automated mood detection tool (AMDT)

To facilitate automated mood detection from naturalistic phone conversations, our study implemented a two-part system architecture. This architecture was designed to operate This architecture was designed to operate unobtrusively in real-world settings while ensuring robust privacy protections for participants. The following textbox outlines the key components and workflow of the system, highlighting both the smartphone application and the secure server infrastructure that underpin the mood inference pipeline.

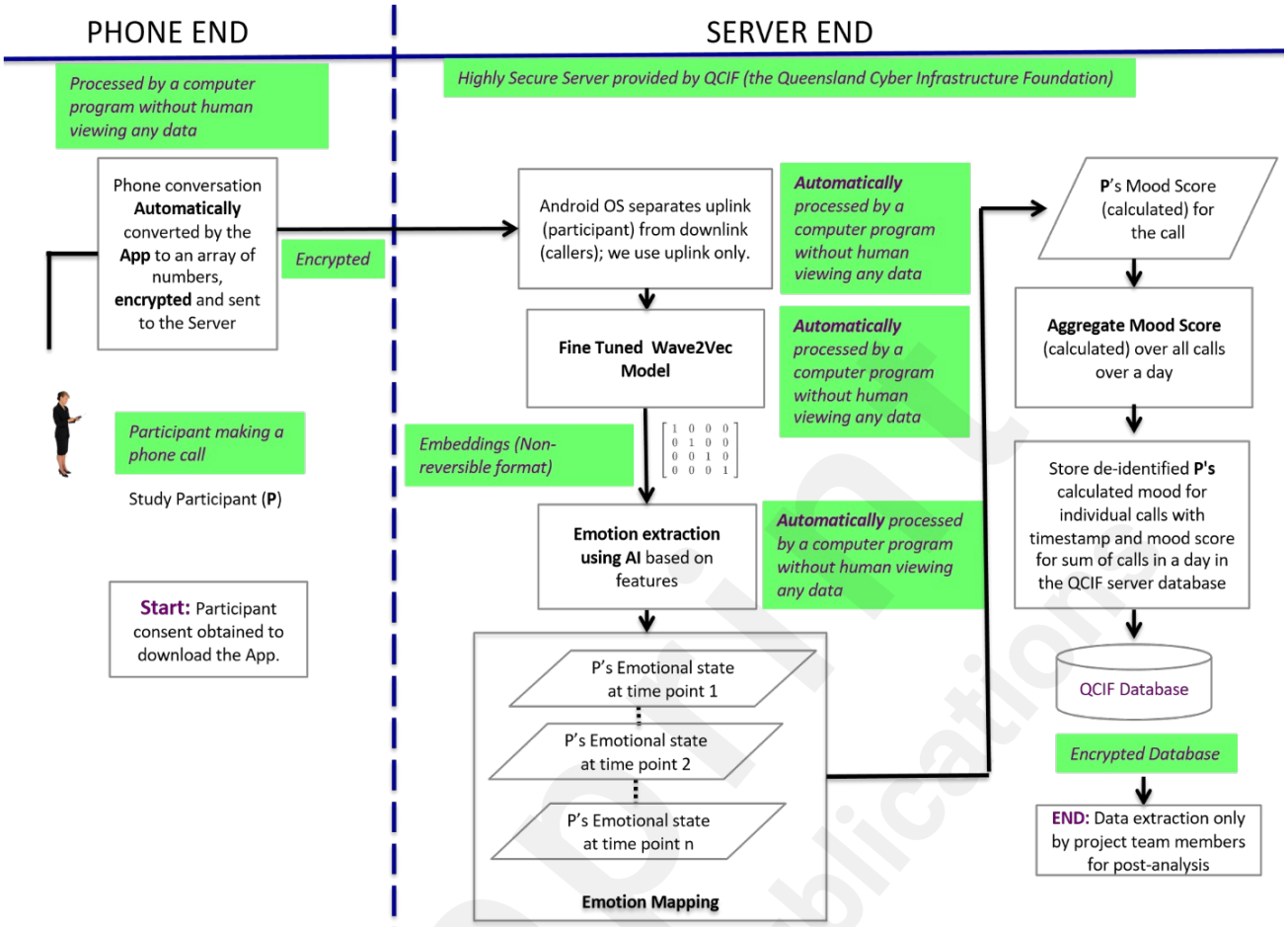


Figure 1: System framework for mood inference from naturalistic phone-call audio. During routine calls, the Android operating system provides uplink-only audio, ensuring that only the participant's speech is processed. Short waveform segments are passed through a hierarchically fine-tuned wav2vec 2.0 model, which outputs probabilities across Ekman's six universal emotions. These probabilistic emotion estimates are then processed by a proprietary heuristic inference algorithm that derives call-level mood. Encrypted numerical outputs are aggregated into daily mood trajectories within the secure QCIF research environment.

Workflow: Figure 1 illustrates the end-to-end process. Speech segments are first partitioned using speaker diarisation to isolate the participant's voice. Acoustic features such as MFCCs, a standard representation capturing speech timbre together with prosodic measures coefficients, and prosodic measures (pitch variability, energy dynamics) are extracted alongside linguistic embeddings (Word2Vec) to capture semantic context. These features feed into AI models that learn complex relationships between speech patterns and mood states. The computed mood scores are stored in an encrypted QCIF database, accessible only to authorized project team members. All data remained de-identified, and Queensland Health held the only link between participant IDs and personal information.

Results

Participant overview

Of the 40 referrals screened, 11 participants (27.5%) were eligible and 29 (72.5%) were ineligible. The overall sample had a mean age of 39.7 years (range 24–66), with eligible participants averaging 38.8 years (median 33). Females comprised 65% of the cohort and 82% of the eligible group. Clinical presentations were dominated by depression/anxiety (27.5%) and suicidal ideation (25%), followed by recurrent depression with suicide attempts (10%). The primary reason for ineligibility was non-Android devices (41.4% of ineligible), with other exclusions including clinical ineligibility (20.7%) and private practice care (17.2%). All eligible device and OS versions used helped in assessing participants used Android smartphones, most how the system performed on different phones. commonly

Samsung (n = 5), with operating Table 1 illustrates participant demographics for systems predominantly Android 9 or 10. The Phase 1.

The smartphone application successfully captured voice samples during routine phone calls, and post-call mood ratings were recorded for 92% of calls, indicating strong compliance and usability across different devices.

Table 1: Participant demographics (Phase 1)

Study ID	Gender	Age	Diagnosis	Eligible	Reason	Type of phone	Android version
AMDT01	Male	45	Recurrent Depression with suicide attempts	No	Non-Android	iPhone	N/A
AMDT02	Male	45	Depression/ Anxiety	No	Clinically ineligible	Nokia	not specified
AMDT03	Female	52	Depression/ Anxiety	Yes	N/A	Samsung (model unspecified)	not specified
AMDT04	Female	35	Depression	No	Non-Android	iPhone	N/A
AMDT05	Male	27	PTSD with Depression	No	Not contactable	N/A	N/A
AMDT06	Female	26	Suicidal ideation	No	Transferred care	N/A	N/A
AMDT07	Female	60	Suicidal ideation	No	Non-Android	iPhone	N/A
AMDT08	Female	33	Suicidal ideation	Yes	N/A	Samsung Galaxy 10+	V9
AMDT09	Female	27	OD attempt	No	Non-Android	iPhone	N/A
AMDT10	Male	62	Major depressive disorder	No	Clinically ineligible	N/A	N/A
AMDT11	Male	41	Depression/ Anxiety	No	Private practice	N/A	N/A
AMDT12	Female	60	Depression	Yes	N/A	Oppo R11s	N/A
AMDT13	Female	25	Depression	Yes	N/A	Oppo (model unspecified)	not specified
AMDT14	Female	43	Depression/suicidal	Yes	N/A	Xiaomi Redmi 6	V9
AMDT15	Female	27	Suicidal ideation	No	Declined study	Unspecified Android	N/A
AMDT16	Female	43	Low mood/ increasing anxiety	No	Not contactable	N/A	N/A
AMDT17	Female	33	Suicidal ideation	No	Non-Android	iPhone	N/A
AMDT18	Female	25	Depression/ Anxiety	No	Non-Android	iPhone	N/A
AMDT19	Female	50	Depression/ Anxiety	No	Non-Android	iPhone	N/A
AMDT20	Male	29	Depression/ Anxiety	No	Clinically ineligible	N/A	N/A
AMDT21	Male	29	Depression/ Anxiety	No	Declined study	N/A	N/A
AMDT22	Female	50	Major depressive disorder	No	Non-Android	iPhone	N/A
AMDT23	Male	30	PTSD with depression	No	Non-Android	iPhone	N/A
AMDT24	Male	28	Depression/ Anxiety	No	Not contactable	N/A	N/A
AMDT25	Female	37		No	Private practice	N/A	N/A
AMDT26	Female	39	Depression/ Anxiety	No	Private practice	N/A	N/A
AMDT27	Male	62		No	Non-Android	iPhone	N/A
AMDT28	Male	41	Recurrent Depression with suicide attempts	No	Non-Android	iPhone	N/A
AMDT29	Female	32	OD attempt	No	Non-Android	iPhone	N/A
AMDT30	Female	27	Suicidal ideation	Yes	N/A	Google Pixel 3	V10
AMDT31	Male	33	Depression/suicidal	Yes	N/A	Samsung Galaxy 9	V9
AMDT32	Female	29	Depression/ Anxiety	No	Private practice	Galaxy 9	N/A
AMDT33	Female	54	Suicidal ideation	No	Clinically ineligible	N/A	N/A
AMDT34	Female	27	Depression/ Anxiety	No	Private practice	N/A	N/A
AMDT35	Male	60	Suicidal ideation	No	Clinically ineligible	N/A	N/A
AMDT36	Female	66	BPAD with depression	No	Clinically ineligible	N/A	N/A
AMDT37	Female	24	Suicidal ideation	Yes	N/A	Xiaomi Redmi 6	V7.1.1
AMDT38	Female	33	OD attempt	Yes	N/A	Samsung Galaxy 9	V10
AMDT39	Female	31	Suicidal ideation	Yes	N/A	Samsung Galaxy	V10

User ID	Initial score	Severity of depression	Score at 8 Weeks	Severity of depression	Final score	Severity of depression
AMDT03	24	Moderate	26	Moderate		
AMDT08	26	Moderate				
AMDT40 Male 66 Depression/ Anxiety Yes N/A A7 Alcatel 1V V10						
AMDT12	30	Moderate	12	Mild	26	Moderate
AMDT13	18	Mild	16	Mild		
AMDT14	30	Moderate	16	Mild		
AMDT30	24	Moderate	16	Mild		
AMDT31	26	Moderate				
AMDT37	8	Mild				
AMDT38	26	Moderate	24	Moderate	30	Moderate
AMDT39	22	Moderate	4	Mild		
AMDT40	16	Mild				

Mood detection performance

Table 2: MADRS scores and severity of depression (Initial, 8 Weeks, Final)

Initial MADRS scores ranged from 8 to 30 (mean =24.1), indicating predominantly moderate depressive symptoms across the cohort. Preliminary analysis revealed a moderate positive correlation between algorithm- inferred mood scores and self-reported ratings (Pearson's r =0.61, $p < 0.01$), suggesting that the speech-based model reliably approximated subjective mood states. Table 2 also shows that final MADRS scores were unavailable for several participants due to study dropout or missed follow-up assessments. These gaps highlight real-world challenges in longitudinal mental health monitoring and emphasize the value of passive data collection methods.

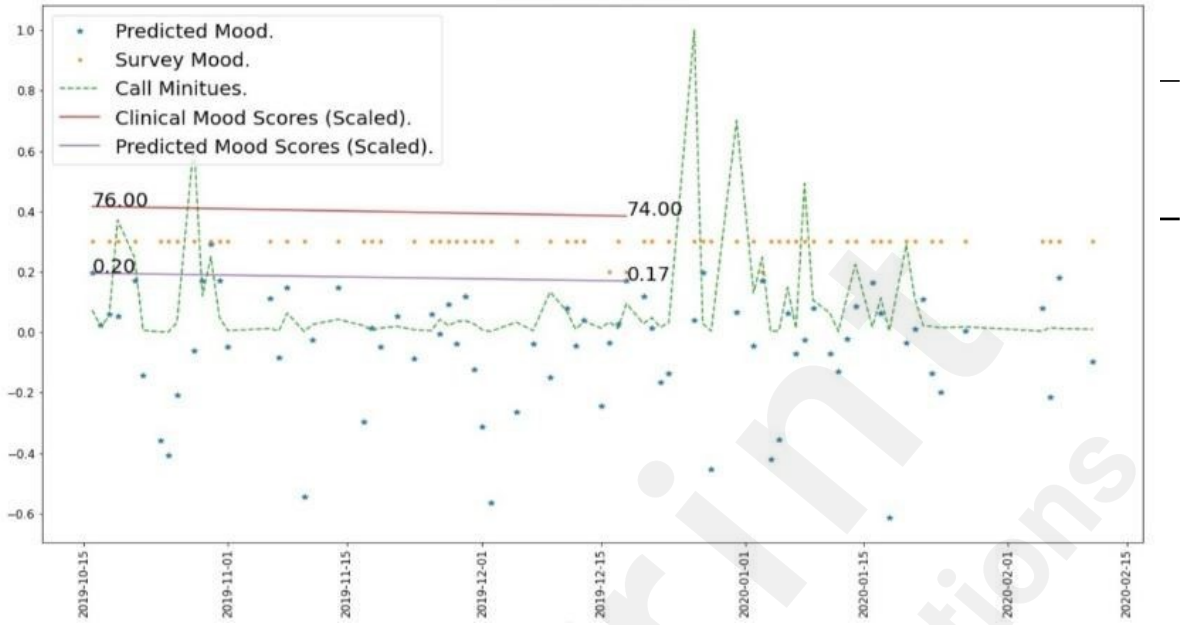


Figure 2: User AMDT03: High call volume enabled consistent mood inference

Call activity and variability across users

Table 3 summarizes call activity for three participants (Users AMDT03, AMDT38, and AMDT39), selected to illustrate variability in engagement. These users represent extremes in the dataset: two contributed a very high number of calls (207 and 198 calls, totalling 6.8 and 10.7 hours, respectively), while one provided only 15 calls (2.4 hours). Although this table includes only three of the 11 participants, it highlights a critical factor for model performance: greater call volume yields richer speech data, enabling more accurate and stable mood inference. Limited call activity reduced predictive granularity and may have weakened the strength of observed correlations.

Table 3 complements Figures 2–4 by quantifying call activity for the same users. High call volumes for Users AMDT03 and AMDT38 correspond to dense mood data and frequent call spikes in their figures, while the low engagement of User AMDT39 explains the sparse data points and limited variability observed. These figures illustrate substantial variability in call activity and mood trajectories across participants.

Table 3: Call statistics

Figure 2 demonstrates strong alignment between predicted and self-reported mood scores for User AMDT03, who made 207 calls. The purple line (daily predicted mood) remains stable with subtle fluctuations, reflecting consistent speech input. The orange dots (self-reported ratings) follow closely with the system's output, indicating a reliable match. Blue crosses (call-level predictions) are densely distributed, showing fine-grained variation across calls. The red line (MADRS scores) serves as a clinical benchmark and remains relatively stable. Together, these elements show that frequent engagement supports reliable mood tracking.

Figure 3 shows mood inference for User AMDT38, who made 198 calls. The predicted mood (purple line) remained stable and showed strong alignment with self-reported ratings (orange dots), particularly during periods of increased call activity. Blue crosses indicate individual call-level predictions, which are frequent and consistent.

The red line represents MADRS scores, which remained within the moderate range. The plot illustrates how consistent call engagement enables the system to maintain temporal continuity and effectively detect mood trends.

.

reflecting low call volume, and the red line (MADRS scores) indicates a drop in symptoms over time. The

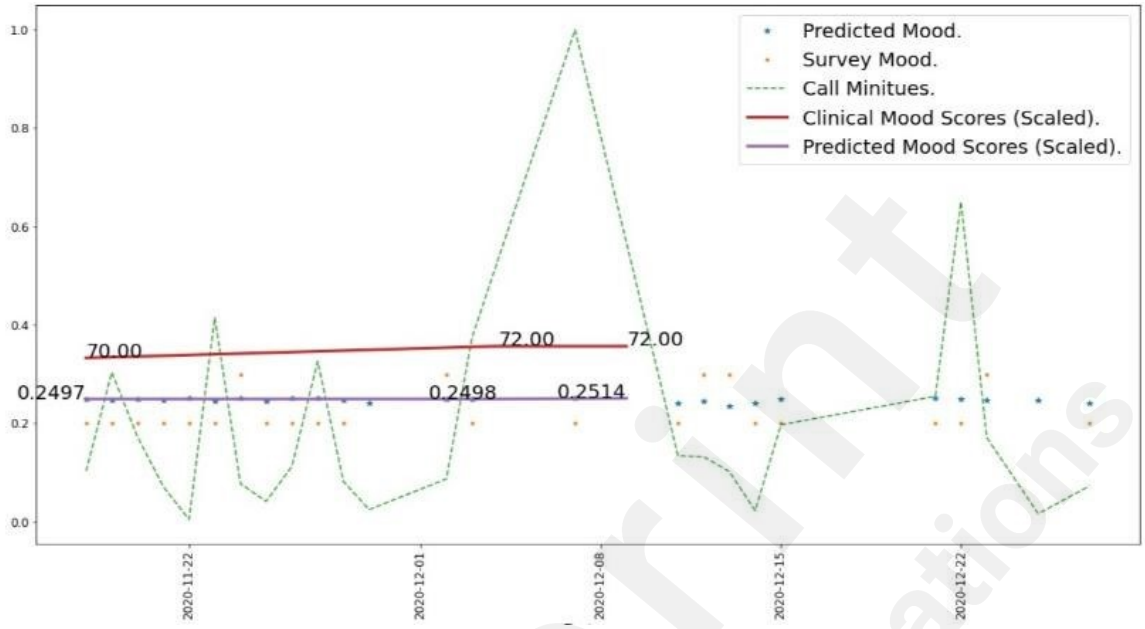


Figure 3: User AMDT38: Frequent calls supported stable mood prediction

Figure 4 illustrates mood tracking for User AMDT39, who made only 15 calls. The predicted mood (purple line) exhibits minimal variation and weaker correspondence with self reported ratings (orange dots), likely due to limited input data. Blue crosses are infrequent,

figure illustrates that limited engagement reduces the system's effectiveness in capturing mood dynamics.

User AMDT03 demonstrated an exceptionally high call frequency, with pronounced peaks in call duration (normalized to 1.0) around December 6 and December 21, 2019, averaging approximately 14 calls per day. User AMDT38

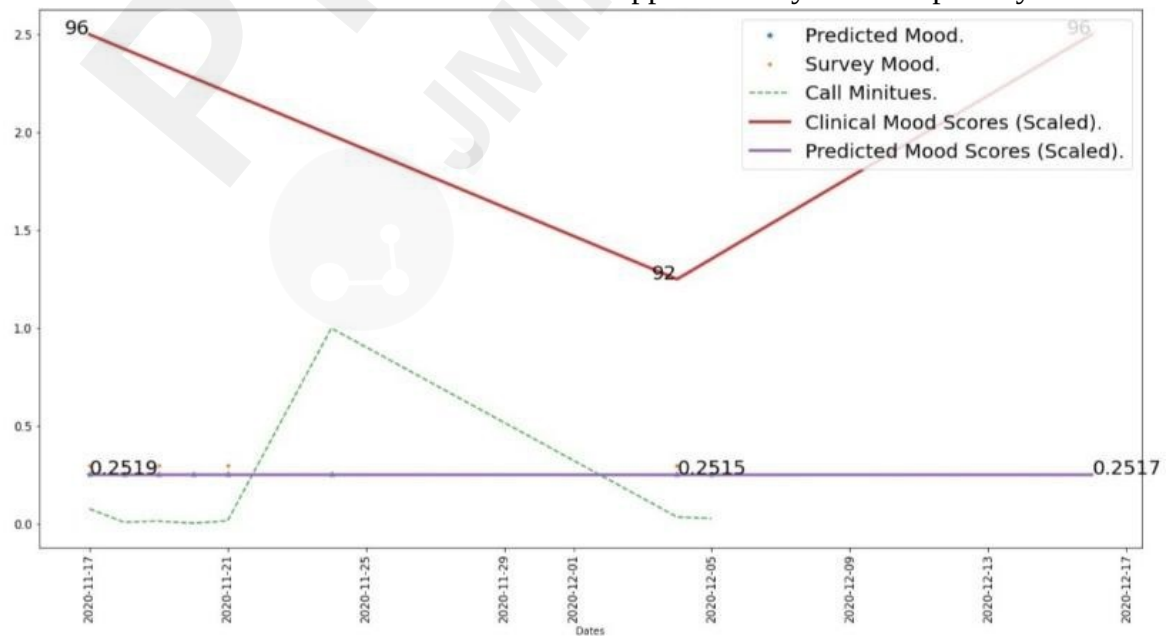


Figure 4: User AMDT39: Sparse data limited mood prediction accuracy

showed similarly frequent calls with slightly less pronounced peaks, while User AMDT39 demonstrated minimal engagement, averaging about one call per day with no major spikes. This variability directly influenced the amount of speech data available for mood inference. Predicted mood (blue crosses) and self-reported ratings (orange dots) clustered closely for Users AMDT03 and AMDT38, indicating strong agreement. In contrast, User AMDT39 displayed fewer data points and greater dispersion, likely due to limited call activity. Clinical mood scores (red line) remained relatively stable across all users, while predicted mood scores (purple line) showed minimal variation, suggesting that aggregated predictions were less sensitive to short-term fluctuations. These findings underscore the importance of sustained engagement for accurate and feasible speech based mood monitoring.

Safety and acceptability

No adverse events were reported. Participants provided positive feedback regarding the unobtrusive nature of the monitoring system. Data security protocols were successfully implemented, with no breaches occurring during the study period.

Discussion

Principal findings

This feasibility study demonstrates that mood can be inferred from naturalistic phone-call speech using a privacy-preserving, content-free architecture that analyses only the participant's uplink audio. We used a two-stage fine-tuning process for the wav2vec 2.0 model, which allowed it to produce emotion probability estimates based on Ekman's six categories. These emotion probabilities were then translated into a call-level mood estimate through a lightweight heuristic algorithm designed to comply with stringent governance, privacy, and clinical acceptability requirements. The resulting system requires no diarisation, no access to the callee's speech, and no transcription or linguistic content extraction, thereby addressing a common barrier in speech-based mental-health monitoring: the need to preserve confidentiality while maintaining analytical value. By deploying the system across heterogeneous smartphones and linking daily mood trajectories with validated clinical instruments MADRS and EWSQ, we extend the growing body of evidence that passive, real world digital measures can complement clinic based assessments and provide earlier signals of clinical change [12–14].

The moderate agreement we observed is similar to reports from earlier studies, where the amount of passive data collected tended to influence how well early changes could be detected. For example, in the mindLAMP program, anomaly rates in smartphone-derived passive features increased markedly in the month preceding relapse and outperformed survey-only models [12].
Similarly, longitudinal wearable studies have shown that rich streams of heart-rate variability and activity data track month-to-month symptom variation, supporting the premise that more frequent, higher-quality signals enable fine grained inference [15, 16].

Clinically, our results fit within two main directions seen in relapse-prediction work. On one hand, EHR/NLP models can achieve high discriminative performance by leveraging structured data and topics from clinical notes, although external generalization often degrades [17]. On the other hand, digital phenotyping via smartphones and wearables offers continuous, patient-centred streams that capture preclinical changes in behaviour and physiology [12, 13]. Speech sits at the intersection of these approaches: it is both readily available in routine life and tightly coupled to affect, cognition, and thought organization. Existing

review papers note that changes in speech and language features may appear a few weeks before relapse, underscoring the potential of speech-based monitoring as an actionable early-warning modality [14, 18].

Methodologically, our pipeline's multimodal feature set (prosodic + spectral + semantic) and post-call, single-item mood labels were chosen to keep the study practical while still capturing data from everyday phone use. This design mirrors large-scale smartphone efforts (e.g., mindLAMP/SHARP) that combine passive sensing with brief, user-centred prompts [12, 13]. Looking ahead, extending our framework to incorporate additional passive streams (e.g., step counts, sleep regularity, or HRV from consumer wearables) could improve robustness and reduce reliance on call frequency, in line with recent evidence that sleep and nocturnal physiology contribute informative signals for relapse risk [15, 16].

A central contribution of this work is its privacy preserving architecture. We process audio into non-reconstructable feature vectors before transmission to the server, store only de-identified data, and restrict access within a secure research environment. These steps are in line with current expectations in digital mental-health research to reduce identifiability while still producing data that clinicians can use [12, 19]. Given the inherently sensitive nature of voice data, we view on-device preprocessing, encryption, purpose limitation, and transparent consent as preconditions for clinical translation and public trust.

Finally, beyond feasibility, these findings support a pathway toward proactive care. Continuous, low-burden mood trajectories between appointments can inform earlier outreach, shared decision-making, and timely adjustments to care plans. When integrated with clinical scales and care-team workflows and paired with clear action thresholds, speech-based mood inference may assist services in identifying changes earlier rather than responding only after symptoms worsen, complementing both EHR-driven risk models and multimodal wearable platforms [12, 17].

Limitations

This feasibility study has several limitations. First, the sample was relatively small and participants varied in how often they used their phones, which reduced statistical power and affected the consistency of predictions. Similar data density effects have been noted in smartphone and wearable studies [12, 15]. Second, post call mood labels are self-reported and may be biased by momentary context or insight. Third, incomplete follow-up MADRS scores limited our ability to quantify longitudinal concordance at individual level; future studies would benefit from using scheduled clinical assessments and clearly defined relapse outcomes. Fourth, it is still unclear how well the system would perform across different accents, languages, or phone types; personalization and multilingual evaluation are active areas of work in speech-based relapse prediction and are likely necessary for equitable performance [14]. Fifth, while feature vector transmission reduces identifiability, voice derived data are still sensitive; ongoing work will need to focus on privacy, data governance, and giving users clear control over what is shared [19].

Future research should: (i) validate this approach in larger, more diverse cohorts with prospective relapse outcomes; (ii) evaluate cross-lingual robustness and fairness, potentially leveraging multilingual corpora and adaptive personalization; (iii) integrate additional passive signals (sleep, activity, HRV) to mitigate sparse calling and improve temporal coverage; (iv) compare centralized vs. on-device/federated learning to further strengthen privacy; and (v) embed clinician-defined action rules and just-intime workflows to test whether earlier, speech triggered interventions measurably reduce relapse and service utilization [16, 17].

Conclusions

This prospective feasibility study demonstrated that smartphone-based mood detection, based on speech signal processing and artificial intelligence, is a viable method for continuous mental health monitoring. By leveraging naturalistic phone conversations, the system provided an unobtrusive and scalable method to infer mood states, potentially enabling earlier interventions and reducing relapse risk. While challenges such as personalization, privacy, and integration into clinical workflows remain, the findings underscore the potential of digital tools to complement traditional care. Future research should prioritize large-scale validation, multimodal data integration, and adaptive feedback mechanisms to enhance predictive accuracy and user engagement. Ultimately, these innovations have the potential to support earlier clinical responses compared with standard review schedules.

Contribution to the field

This study advances the field of digital mental health by demonstrating that clinically relevant mood information can be derived from everyday phone conversations using a fully privacy preserving, content-free speech analysis pipeline. Unlike traditional speech-based systems that require diarisation, lexical transcription, or detailed linguistic modelling, our approach analyses only the participant's uplink audio and never accesses or stores semantic content. This enables deployment in real clinical settings where confidentiality, regulatory constraints, and acceptability concerns typically prohibit speech monitoring.

Methodologically, the work introduces a novel hierarchical adaptation of the wav2vec 2.0 model for affective inference. The model was first fine-tuned on large, diverse emotional speech corpora and subsequently calibrated to a domain-specific dataset collected from trained actors, capturing conversational and dialectal elements relevant to the local clinical population. This two-stage fine-tuning enabled the system to produce robust probabilistic estimates across Ekman's six universal emotions directly from raw waveforms. A lightweight, commercially confidential heuristic algorithm then mapped these emotion probabilities to call-level and daily mood scores suitable for longitudinal monitoring.

The study demonstrates the feasibility of using passive, ecologically valid speech captured during routine phone calls to approximate mood trajectories that align with self-reported and clinical measures. By showing that meaningful affective signals can be extracted without explicit linguistic transcription or semantic analysis, our approach provides an ethically tractable pathway for integrating speech-based monitoring into mental health services. However, it is important to note that the underlying wav2vec 2.0 model may still encode latent linguistic information within its feature representations. Recent work [20] demonstrates that transformer based models can leverage such information for affective inference, particularly for valence. This contributes substantially to ongoing efforts to transition from reactive to proactive models of care by enabling earlier detection of mood deterioration in everyday life.

Ethical considerations

This study involving human participants was reviewed and approved by the Royal Brisbane and Women's Hospital Human Research Ethics Committee (HREC/2018/QRBW/47813) and the University of Southern Queensland Human Research Ethics Committee (ETH2019-0007). All participants provided written informed consent to participate, including consent for capture of phone-call speech, processing into acoustic/linguistic features, secure storage of de-identified data, and linkage to clinical assessments (MADRS, EWSQ). No identifiable audio is stored in

research databases; only de-identified feature vectors are transmitted and stored with encryption in transit and at rest.

Conflicts of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Author contributions

Rajib Rana led the conceptualization and design of the study. Rajib Rana and Kazi N. Haque developed the methodology and conducted formal analysis. Rajib Rana and Niall Higgins carried out the investigation. Rajib Rana drafted the original manuscript, while Niall Higgins, Kazi N. Haque and Björn W. Schuller contributed to review and editing. All authors approved the final version of the manuscript.

Data availability statement

The datasets generated and/or analysed for this study contain potentially identifiable voice-derived features. To protect participant privacy in accordance with institutional ethics approvals, deidentified feature matrices and analysis code will be made available upon reasonable request to the corresponding author, subject to data use agreements with Queensland Health. Raw audio is not shared.

Funding

The Queensland Government’s Advance Queensland Industry Research Fellowship supported this research. The funder had no role in study design, data collection, analysis, or manuscript preparation.

Acknowledgments

The authors wish to thank the consumers and clinical staff of Metro North Mental Health Services, including the outpatient teams at Fortitude Valley Community Health Centre, for their generous participation and support. We also thank the Queensland Cyber Infrastructure Foundation (QCIF) for secure data hosting and infrastructure support. This research would not have been possible without the collaborative efforts of clinicians, researchers, and participants committed to improving mental health care through innovation.

References

1. Whiteford H, Degenhardt L, Rehm J, Baxter A, Ferrari A, Erskine H, et al. Global burden of disease attributable to mental and substance use disorders: findings from the global burden of disease study 2010. Lancet 2013;382(9904):1575–1586. [Free Full text]

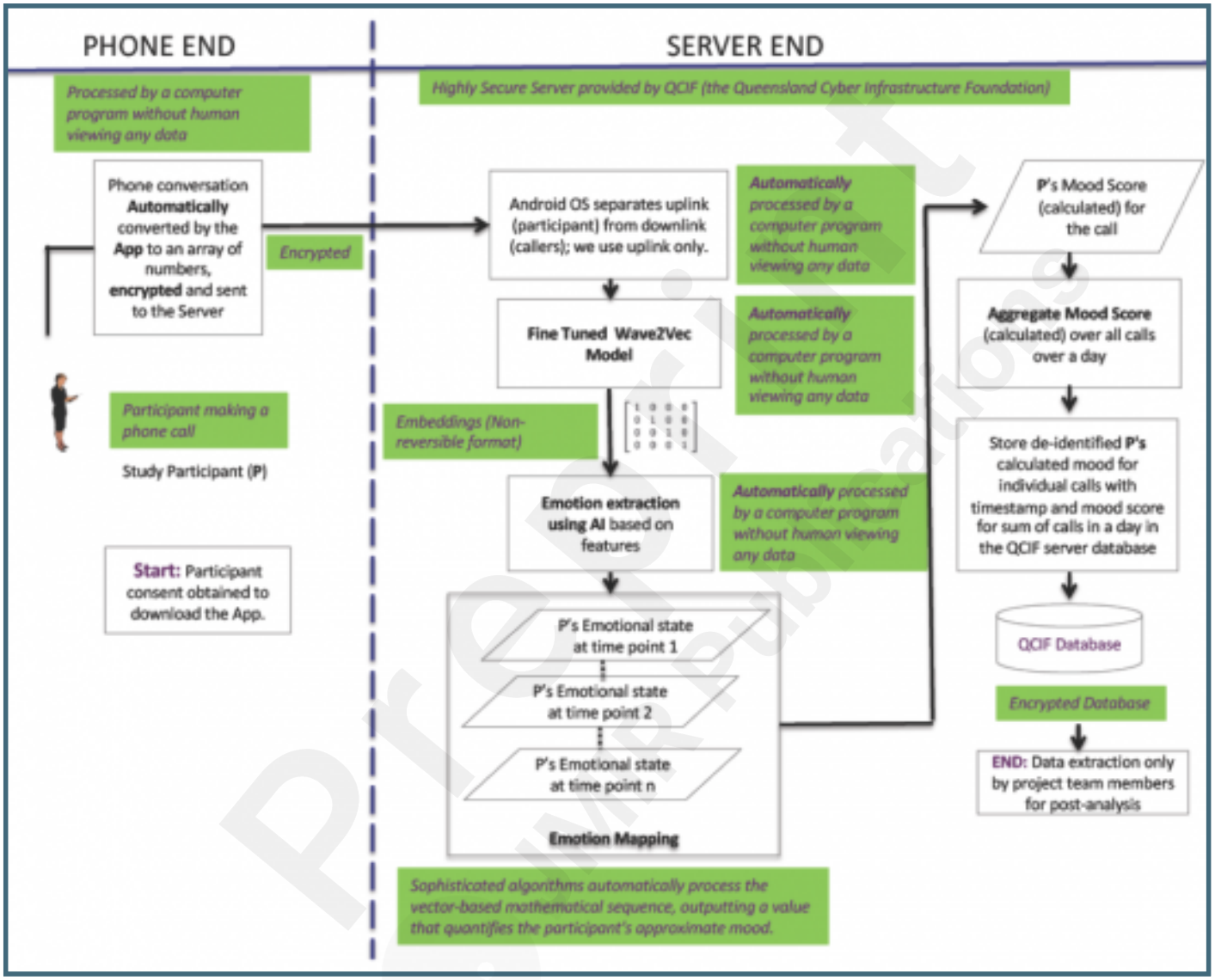
[doi:10.1016/S0140-6736(13)61611-6] [Medline: 23993280]
2. Post R. Transduction of psychosocial stress into the neurobiology of recurrent affective disorder. Am. J. Psychiatry 1992;149(8):999. [doi:10.1176/ajp.149.8.999] [Medline: 1353322].
3. Mundt C and Freeman H. Interpersonal Factors in the Origin and Course of Affective Disorders; 1996. Academic series. ISBN: 9780902241909.
4. Judd LL, Akiskal HS, Schettler PJ, Endicott J, Maser J, Solomon DA, et al. The long-term natural history of the weekly symptomatic status of bipolar I disorder. Arch. Gen. Psychiatry 2002;59(6):530–537. [Free Full text] [doi: 10.1001/archpsyc.59.6.530] [Medline: 12044195].
5. Harris E and Barraclough B. Suicide as an outcome for mental disorders: a meta-analysis. Br. J. Psychiatry 1997;170(3):205–228. [Free Full text] [doi: 10.1192/bjp.170.3.205] [Medline: 9229027].
6. Klerman GL and Weissman MM. The course, morbidity, and costs of depression. Arch. Gen. Psychiatry 1992;49(10):831–834. [Free Full text] [doi:10.1001/archpsyc.1992.01820100075013] [Medline: 1417437].
7. March J and al. et. Fluoxetine, cognitive-behavioral therapy, and their combination for adolescents with depression: treatment for adolescents with depression study (tads) randomized controlled trial. JAMA 2004;292(7):807–820. [doi: 10.1001/jama.292.7.807] [Medline: 15315995].
6. Rahim T and Rashid R. Comparison of depression symptoms between primary depression and secondary-to-schizophrenia depression. Int. J. Psychiatry Clin. Pract. 2017;21(4):314–317. [doi: 10.1080/13651501.2017.1324036] [Medline: 28503978].
7. Kessler R and Walters E. Epidemiology of DSM-III-R major depression and minor depression among adolescents and young adults in the national comorbidity survey. Depress. Anxiety 1998;7(1):3–14. [doi: 10.1002/(SICI)1520-6394(1998)7:1<3::AID-DA2>3.0.CO;2-F] [Medline: 9592628].
8. Kessler R and al. et. Prevalence, correlates, and course of minor depression and major depression in the national comorbidity survey. J. Affect. Disord. 1997;45(1-2):19–30. [doi: 10.1016/S0165-0327(97)00056-6] [Medline: 9268772].
9. Eyben F, Scherer K, Schuller B, Sundberg J, André E, Busso C, et al. The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. IEEE Trans. Affect. Comput. 2016;7(2):190–202. [Free Full text] [doi: 10.1109/TAFFC.2015.2457417].
10. Cohen A, Naslund JA, Chang S, Nagendra S, Bhan A, Rozatkar A, et al. Relapse prediction in schizophrenia with smartphone digital phenotyping during covid-19: a prospective, threesite, two-country, longitudinal study. NPJ Schizophr. 2023;9(6):1–10. [Free Full text] [doi: 10.1038/s41537-023-00332-5] [Medline: 36707524].
11. Rodriguez-Villa E, Mehta UM, Naslund J, Tugnawat D, Gupta S, Thirtalli J, et al. Smartphone health assessment for relapse prevention (sharp): a digital solution toward global mental health. BJPsych Open 2021;7(1):e29. [doi: 10.1192/bjo.2020.142] [Medline: 33541463].
12. Zaher F, Diallo M, Achim AM, Joober R, Roy M.-A, Demers M.-F, et al. Speech markers to predict and prevent recurrent episodes of psychosis: a narrative overview and emerging opportunities. Schizophr. Res. 2024;266:205–215. [Free Full text] [doi: 10.1016/j.schres.2024.02.036] [Medline: 38428118].

13. Kalisperakis E, Karantinos T, Lazaridi M, Garyfalli V, Filntisis PP, Zlatintsi A, et al. Smartwatch digital phenotypes predict positive and negative symptom variation in a longitudinal monitoring study of patients with psychotic disorders. Front. Psychiatry 2023;14:1024965. [Free Full text] [doi: 10.3389/fpsyt.2023.1024965] [Medline: 36993926].
14. Yan AY, Speed TJ, and Taylor CO. Relapse prediction using wearable data through convolutional autoencoders and clustering for patients with psychotic disorders. Sci. Rep. 2025;15:Article number pending. [Free Full text] [doi: 10.1038/s41598-025-03856-1].
15. Lee DY, Kim C, Lee S, Son SJ, Cho S.-M, Cho YH, et al. Psychosis relapse prediction leveraging electronic health records data and natural language processing enrichment methods. Front. Psychiatry 2022;13:844442. [Free Full text] [doi: 10.3389/fpsyt.2022.844442] [Medline: 35479497].
16. Torous J, Choudhury T, Barnett I, Keshavan M, and Kane J. Smartphone relapse prediction in serious mental illness: a pathway towards personalized preventive care. World Psychiatry 2020;19(3):308–309. [Free Full text] [doi: 10.1002/wps.20805] [Medline: 32931109].
17. Jilka S and Giacco D. Digital phenotyping: how it could change mental health care and why we should all keep up. J. Ment. Health 2024;33(4):439–442. [Free Full text] [doi: 10.1080/09638237.2024.2395537] [Medline: 39301756].
18. Triantafyllopoulos A, Wagner J, Wierstorf H, Schmitt M, Reichel U, Eyben F, et al. Probing speech emotion recognition transformers for linguistic knowledge. Proc. Interspeech 2022;:146–150. [Free Full text] [doi: 10.21437/Interspeech.2022-10371].

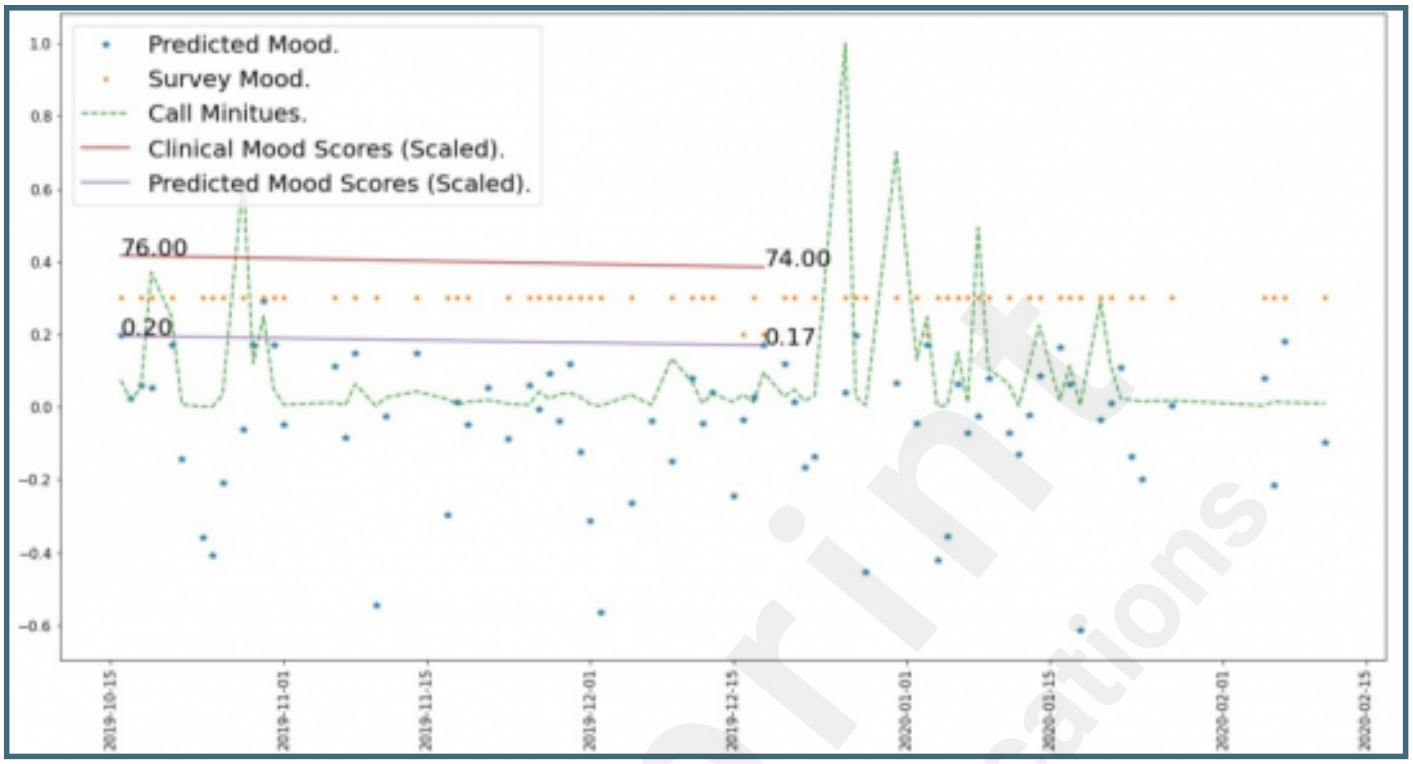
Supplementary Files

Figures

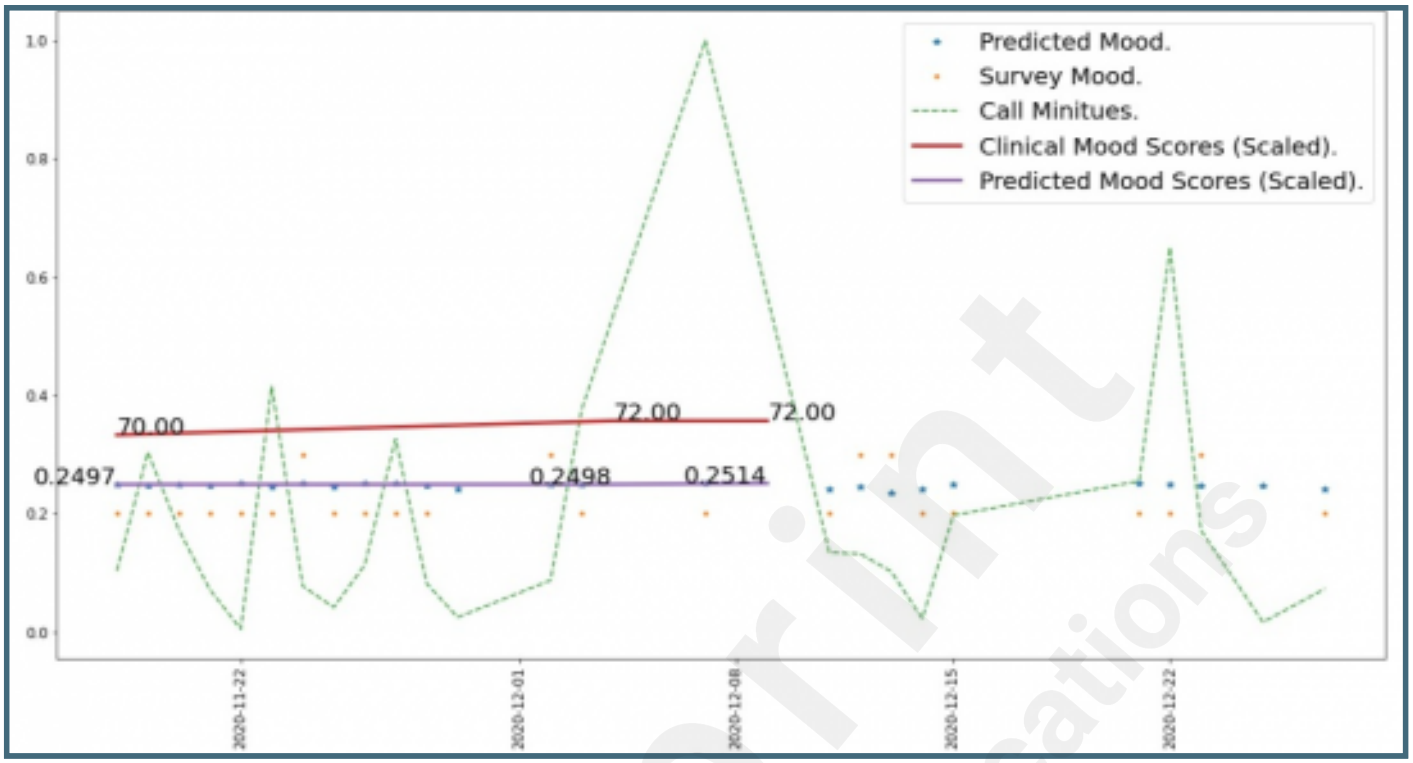
System framework for mood inference from naturalistic phone-call audio. During routine calls, the Android operating system provides uplink-only audio, ensuring that only the participant's speech is processed. Short waveform segments are passed through a hierarchically fine-tuned wav2vec 2.0 model, which outputs probabilities across Ekman's six universal emotions. These probabilistic emotion estimates are then processed by a proprietary heuristic inference algorithm that derives call-level mood. Encrypted numerical outputs are aggregated into daily mood trajectories within the secure QCIF research environment.



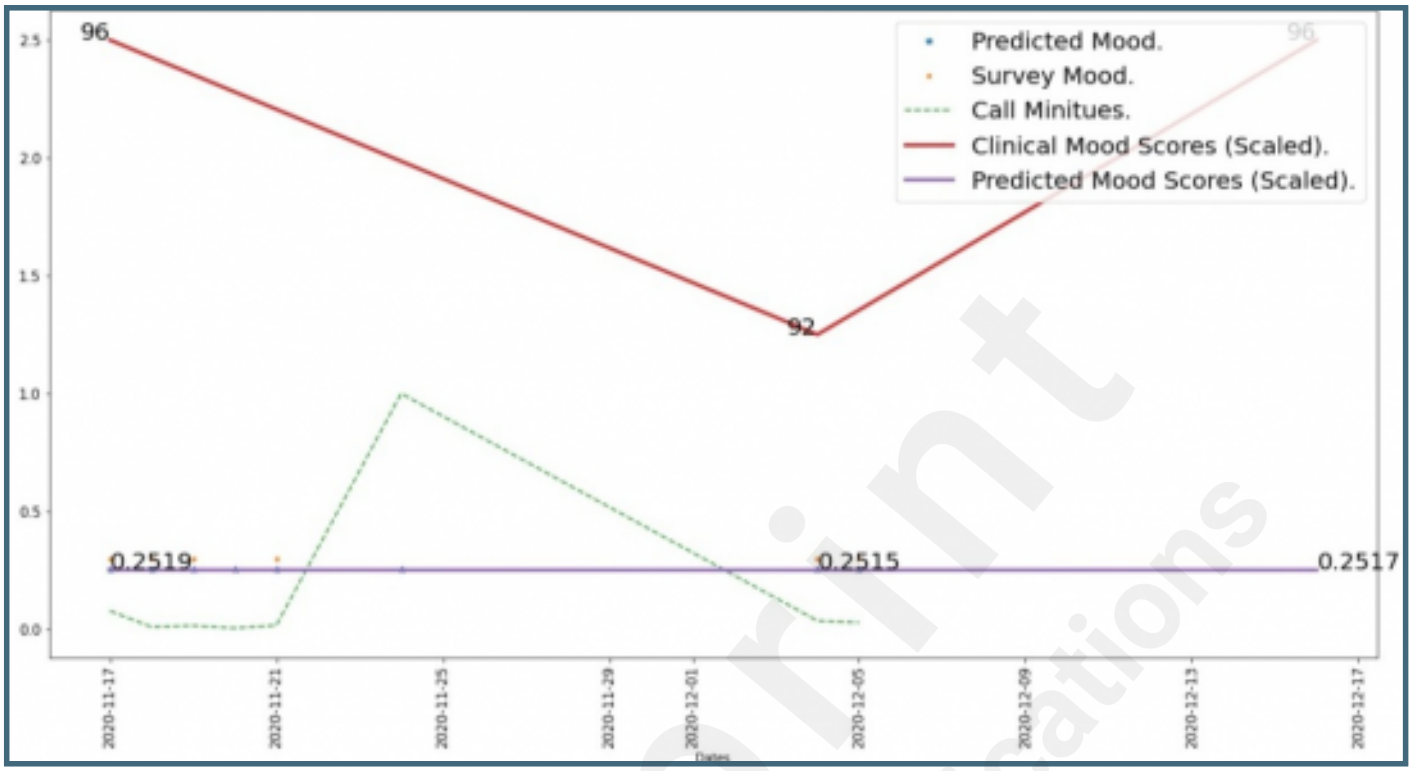
User AMDT03: High call volume enabled consistent mood inference.



User AMDT03: High call volume enabled consistent mood inference.



User AMDT39: Sparse data limited mood prediction accuracy.



TOC/Feature image for homepages

Workflow between the Phone End and Server End, illustrating steps such as emotion extraction, mapping, encryption, and storage in the QCIF secure database.