

Towards Interpretable Depression Detection: Linking Acoustic Features to DSM-5 Indicators

Jonas Länzlinger¹, Katharina O.E. Müller², Burkhard Stiller², Bruno Rodrigues¹

¹Embedded Sensing Group ESG, School of Computer Science SCS, University of St. Gallen HSG, Switzerland

²Communication Systems Group CSG, Department of Informatics IfI, University of Zurich UZH, Switzerland

E-mail:jonas.laenzlinger@student.unisg.ch|stiller,mueller@ifi.uzh.ch|bruno.rodrigues@unisg.ch

Abstract—Depression affects millions worldwide, yet diagnosis relies on subjective self-reports that may miss authentic behavior. This paper presents an approach linking speech acoustics to DSM-5 depressive-behavior indicators through a transparent Linkage Framework. Unlike black-box models, the framework explicitly maps acoustic features (pitch variability, pauses, speech tempo) to clinical indicators, enabling interpretable, indicator-level outputs. The system runs locally on commodity hardware (HW) to preserve privacy. Preliminary evaluation on DAIC-WOZ shows directionally consistent associations between acoustic features and DSM-5 indicators for psychomotor change and concentration difficulty, supporting the design rationale. Future work will validate on longitudinal datasets and extend multimodal integration while maintaining edge constraints.

Index Terms—Depression detection, acoustic biomarkers, digital health, passive sensing, smart home

I. INTRODUCTION

Depression is one of the most prevalent mental health disorders, with rates rising across all age groups [1], [2]. The consequences are especially severe for children and adolescents, where early symptoms are often overlooked [3]. While clinical diagnosis relies on DSM-5 criteria and self-report questionnaires like PHQ-9 [4], [5], these methods are susceptible to recall errors and subjectivity [6]. As access to mental health support remains limited [1], innovative detection methods become critical, yet current systems face two barriers: *interpretability* and *privacy*.

The interpretability gap. Deep learning models achieve high accuracy, but their outputs (*e.g.*, a single probability score) provide no insight into *which symptoms* are elevated or *why*. Clinicians cannot integrate such outputs into workflows requiring symptom-level reasoning aligned with DSM-5 [7]. A transparent mapping between signals and clinical indicators is essential.

The privacy imperative. Speech data is inherently sensitive in household contexts. Cloud offloading raises major risks: recordings reveal identity and health status. Local processing ensures raw audio never leaves the home, which is critical for longitudinal monitoring.

Although combining speech, text, physiology, and behavior yields the most reliable assessments [8], [9], integrating heterogeneous signals on constrained household devices remains impractical. Therefore, our approach concentrates on acoustic data from speech: a modality that carries clinically validated

cues of depression and can be processed transparently and efficiently at the edge.

Related work. Depression detection has been explored across text, speech, vision, and mobile sensing. Text-based methods analyze social media posts using transformer embeddings [10], [11], while vision approaches leverage facial expressivity and gaze [12], [13]. Digital phenotyping uses passively collected signals from phones and wearables [14], [15]. For speech, prior work shows that acoustic cues (reduced pitch variability, slower speech, longer pauses) correlate with depressive states [16], [17]. However, most studies report aggregate accuracy without exposing which feature supports which DSM-5 indicator [18]. Without precise feature-indicator mappings, outputs remain hard to interpret in clinical workflows.

It had been investigated whether acoustic features can be mapped to DSM-5 indicators in a transparent, resource-aware way. The challenge is that acoustic cues exhibit many-to-many relations with symptoms and are sensitive to confounders (age, sex, channel conditions), requiring a structured mapping that remains efficient for edge execution while producing interpretable outputs.

This paper presents an approach that links acoustic features to DSM-5 [4] depression indicators through a transparent specification that runs locally, preserving privacy while enabling near-real-time analysis. The contributions are:

- A **Linkage Framework** that maps acoustic features to DSM-5 depressive-behavior indicators through explicit, testable rules.
- An **edge-first design** with a privacy-preserving pipeline demonstrating real-time throughput on commodity HW.
- **Preliminary evaluation** on DAIC-WOZ [19] showing directionally consistent feature–indicator associations.

II. SYSTEM DESIGN

Moving beyond aggregate depression scores to symptom-level understanding is essential for appropriate treatment [7]. While deep learning achieves high accuracy, it often lacks the interpretability needed for clinical workflows [20]. Our approach prioritizes explainability through a structured pipeline (Figure 1).

The architecture comprises: (1) a **data ingestion layer** that captures audio via microphones and applies VAD (SilerovAD [21]) to isolate speech; (2) a **processing layer** extracting

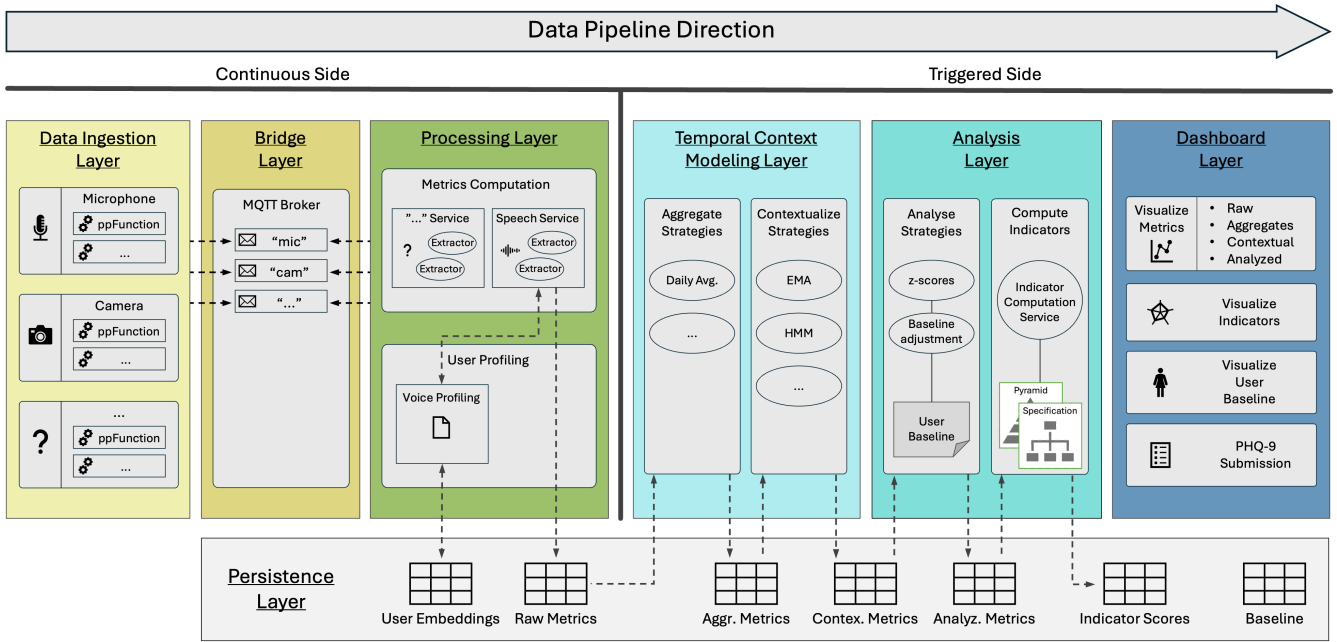


Fig. 1: High-level system architecture showing the data pipeline from audio ingestion to DSM-5 indicator scoring.

low-level descriptors (F_0 , intensity, jitter, shimmer) on 25 ms frames, aggregated into high-level descriptors over 10 s windows; (3) a **temporal context layer** applying indicator-specific smoothing; and (4) an **analysis layer** mapping features to DSM-5 indicators. All processing above ingestion runs on-device, preserving privacy.

A. Linkage Framework (LF)

The LF bridges raw audio with DSM-5 symptom indicators (Table I) through a layered, clinically-grounded mapping. Unlike end-to-end models that learn opaque representations, the LF makes each transformation explicit and testable.

Layer structure. The framework comprises five layers: *measurements* (frame-level F_0 , intensity, jitter, shimmer) → *features* (functionals like F_0 range, harmonics-to-noise ratio, spectral slope) → *biomarkers* (monopitch, monoloudness, slowed speech, prolonged pauses) → *indicators* (DSM-5 symptom items) → *analysis* (counting rule). Each layer has defined inputs/outputs, enabling modularity: features can be extended without changing clinical logic.

Feature categories. Two main feature families are extracted. *Prosodic features* (F_0 statistics, intensity range, speech/articulation rate, pause ratios) capture pitch, loudness, and timing patterns associated with monotony and slowed speech [16]. Depressed individuals often exhibit “prosodic flattening,” characterized by compressed pitch range and diminished loudness dynamics [22]. *Source/quality features* (jitter, shimmer, HNR, CPP) reflect phonatory stability and breathiness, linked to fatigue and emotional flattening. Elevated jitter and shimmer indicate vocal fold tension changes, while reduced HNR suggests breathiness consistent with psychomotor fatigue [22].

TABLE I: DSM-5 symptom indicators for MDD [4].

ID	Symptom Indicator
(1)	Depressed mood
(2)	Loss of interest or pleasure
(5)	Psychomotor retardation or agitation
(8)	Diminished ability to think/concentrate

Note: Full 9 indicators in DSM-5; those most observable from speech.

Clinical grounding. Each feature-indicator link is grounded in clinical literature. Reduced F_0 variability (monopitch) reflects affective flattening in depressed patients, historically characterized as “low, slow, hesitant, and monotonous” speech [22]. Longer pauses and slower articulation rate indicate psychomotor retardation, *i.e.*, difficulty initiating and sustaining speech [16]. Increased jitter and shimmer reflect reduced phonatory control associated with fatigue [17]. These associations are directional: reduced variability increases indicator scores, as does increased pause duration.

Many-to-many mappings. Crucially, mappings are **many-to-many**: one feature like F_0 range contributes to multiple indicators (*e.g.*, depressed mood *and* loss of interest), while each indicator draws from multiple features (Figure 2). This deliberate redundancy increases robustness against noise and missing data. For example, psychomotor retardation (indicator 5) is supported by pause duration, speech rate, articulation rate, and F_0 variability; if one feature is unreliable due to channel noise, others provide compensating evidence.

B. Computing Indicator Scores

The transformation from features to DSM-5 indicator scores follows four steps. First, features are standardized

recall bias. The sample size ($n=64$) limits statistical power; effect sizes show directional consistency but remain moderate after FDR correction, consistent with the multifactorial etiology of depression. The current LF covers four of nine DSM-5 indicators; symptoms like appetite change and sleep disturbance require complementary modalities. Despite these constraints, directional consistency across multiple feature families supports the design rationale.

IV. SUMMARY AND FUTURE WORK

This paper introduced a framework [24] linking acoustic features to DSM-5 indicators through an explicit Linkage Framework. Unlike black-box models, our approach exposes which features support which indicators, enabling symptom-level review. The edge-first design ensures raw speech never leaves the home, addressing privacy concerns for household monitoring.

Preliminary evaluation on DAIC-WOZ shows directionally consistent associations for psychomotor change and concentration difficulty. Streaming evaluation confirms appropriate temporal behavior, damping transients while accumulating sustained patterns.

Future work includes: (1) longitudinal validation in home environments with naturalistic speech; (2) multimodal fusion with wearable signals for indicators not observable from speech; (3) population calibration across languages and demographics; and (4) clinical integration for screening workflows. The open-source implementation enables reproducibility toward transparent mental health monitoring.

REFERENCES

- [1] W. H. Organization, "Depressive disorder (depression)," March 2023, accessed: 29-04-2025. [Online]. Available: <https://www.who.int/news-room/fact-sheets/detail/depression>
- [2] O. Remes, J. F. Mendes, and P. Templeton, "Biological, psychological, and social determinants of depression: A review of recent literature," *Brain Sciences*, vol. 11, no. 12, p. 1633, 2021, pMID: 34942936, PMCID: PMC8699555. [Online]. Available: <https://www.mdpi.com/2076-3425/11/12/1633>
- [3] Agency for Healthcare Research and Quality (US), 2022 *National Healthcare Quality and Disparities Report*. Rockville (MD): Agency for Healthcare Research and Quality (US), October 2022, internet. [Online]. Available: <https://www.ncbi.nlm.nih.gov/books/NBK587174/>
- [4] A. P. Association, *Diagnostic and statistical manual of mental disorders: DSM-5-TR*, fifth edition ed., Washington, DC, 2022.
- [5] K. Kroenke, R. L. Spitzer, and J. B. W. Williams, "The phq-9: validity of a brief depression severity measure," *Journal of General Internal Medicine*, vol. 16, no. 9, pp. 606–613, September 2001.
- [6] R. F. Baumeister, K. D. Vohs, and D. C. Funder, "Psychology as the science of self-reports and finger movements: Whatever happened to actual behavior?" *Perspectives on Psychological Science*, vol. 2, no. 4, pp. 396–403, 2007, pMID: 26151975. [Online]. Available: <https://doi.org/10.1111/j.1745-6916.2007.00051.x>
- [7] I. F. Eiko and R. M. Nesse, "Depression sum-scores don't add up: why analyzing specific depression symptoms is essential," *BMC Medicine*, vol. 13, no. 1, p. 72, 2015. [Online]. Available: <https://doi.org/10.1186/s12916-015-0325-4>
- [8] Z. Zhang, S. Zhang, D. Ni, Z. Wei, K. Yang, S. Jin, G. Huang, Z. Liang, L. Zhang, L. Li *et al.*, "Multimodal sensing for depression risk detection: Integrating audio, video, and text data," *Sensors*, vol. 24, no. 12, p. 3714, 2024.
- [9] R. Jahan and M. M. Tripathi, "Chapter two - multimodal depression detection using machine learning," in *Artificial intelligence, machine learning, and mental health in pandemics*, S. Jain, K. Pandeym, P. Jain, and K. P. Seng, Ed. Academic Press, 2022, pp. 53–72. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/B9780323911962000053>
- [10] A. Sharma and U. Ghose, "Sentimental analysis of twitter data with respect to general elections in india," *Procedia Computer Science*, vol. 173, pp. 325–334, 01 2020.
- [11] F. Cacheda, D. Fernandez, F. J. Novoa, and V. Carneiro, "Early detection of depression: Social network analysis and random forest techniques," *Journal of Medical Internet Research*, vol. 21, no. 6, p. e12554, 2019. [Online]. Available: <https://doi.org/10.2196/12554>
- [12] J. F. Cohn, T. S. Kruez, I. Matthews, Y. Yang, M. H. Nguyen, M. T. Padilla, F. Zhou, and F. De la Torre, "Detecting depression from facial actions and vocal prosody," in *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, 2009, pp. 1–7.
- [13] K. Min, J. Yoon, M. Kang, D. Lee, E. Park, and J. Han, "Detecting depression on video logs using audiovisual features," *Humanities and Social Sciences Communications*, vol. 10, no. 1, p. 788, 2023. [Online]. Available: <https://doi.org/10.1057/s41599-023-02313-6>
- [14] C. Lin, P. Hu, H. Su, S. Li, J. Mei, J. Zhou, H. Leung, "Sensemod: Depression detection on social media," in *Proceedings of the 2020 International Conference on Multimedia Retrieval*, ser. ICMR '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 407–411. [Online]. Available: <https://doi.org/10.1145/3372278.3391932>
- [15] G. Shen, J. Jia, L. Nie, F. Feng, C. Zhang, T. Hu, T.-S. Chua, and W. Zhu, "Depression detection via harvesting social media: A multimodal dictionary learning solution," in *Proceedings of the twenty-sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, 2017, pp. 3838–3844. [Online]. Available: <https://doi.org/10.24963/ijcai.2017/536>
- [16] D. M. Low, K. H. Bentley, S. S. Ghosh, "Automated assessment of psychiatric disorders using speech: A systematic review," *Laryngoscope Investigative Otolaryngology*, vol. 5, no. 1, pp. 96–116, 2020. [Online]. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1002/lto2.354>
- [17] P. Donaghy, E. Ennis, M. Mulvenna, R. R. Bond, N. Kennedy, M. McTear, H. O'Connell, N. Blaylock, R. Brueckner, "A review of studies using machine learning to detect voice biomarkers for depression," *Journal of Technology in Behavioral Science*, vol. —, no. —, pp. 1–15, Dec. 2024. [Online]. Available: <https://doi.org/10.1007/s41347-024-00454-2>
- [18] U. Yadav, A. Sharma, and D. Patil, "Review of automated depression detection: Social posts, audio and video, open challenges and future direction," *Concurrency and Computation: Practice and Experience*, vol. 35, 10 2022.
- [19] U. I. for Creative Technologies, "Daic-woz database & extended daic database," <https://dcapswoz.ict.usc.edu/>, Jun. 2025, accessed: 10-06-2025.
- [20] L.-V. Herm, K. Heinrich, J. Wanner, and C. Janiesch, "Stop ordering machine learning algorithms by their explainability! a user-centered investigation of performance and explainability," *International Journal of Information Management*, vol. 69, p. 102538, 2023. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S026840122200072X>
- [21] S. team, "Silero vad: pre-trained enterprise-grade voice activity detector (vad), number detector and language classifier," <https://github.com/snakers4/silero-vad>, 2024.
- [22] N. Cummins, S. Scherer, J. Krajewski, S. Schnieder, J. Epps, and T. F. Quatieri, "A review of depression and suicide risk assessment using speech analysis," *Speech Communication*, vol. 71, pp. 10–49, 2015. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0167639315000369>
- [23] M. K. Pichora-Fuller and K. Dupuis, "Toronto emotional speech set (TESS)," 2020. [Online]. Available: <https://doi.org/10.5683/SP2/E8H2MF>
- [24] J. Länzlänger, "Depression Detection GitHub Repository," <https://github.com/jonaslanzlanger/depression-detection>, 2025, accessed: 21-06-2025.