Journal of Inclusive Methodology and Technology in Learning and Teaching
ISSN 2785-5104
Anno 6 n. 3 (2026)

EDIZIONI UNIVERSITARIE ROMANE
editoria scientifica dal 1952

Voice, Text, and Large Language Models for Socio-Affective Analysis: Educational Implications and Ethical Considerations

Gianni Fenu [1], Salvatore Carta [1,2], Alessandro Giuliani [1], Alessandro Sebastian Podda [1,*], Marco Manolo Manca [1] and Livio Pompianu [1]

[1] Department of Mathematics and Computer Science, University of Cagliari; {fenu, salvatore, alessandro.giuliani, sebastianpodda, marcom.manca, livio.pompianu}@unica.it
[2] VisioScientiae S.r.l., Cagliari; salvatore.carta@visioscientiae.com
* Correspondence: sebastianpodda@unica.it

Abstract: The increasing use of artificial intelligence in learning analytics is expanding the range of behavioral, linguistic, and socio-affective signals that can be computationally analyzed. This paper investigates the methodological potential and limitations of voice-, text-, and Large Language Model (LLM)-based analysis by reframing an experimental framework originally developed for depression detection as a case study for socio-affective analytics in education. We conducted experiments on the DAIC-WOZ corpus using three analytical approaches: raw-speech modeling with an ECAPA-TDNN architecture and Non-Uniform Speaker Disentanglement; text-based prediction using Word2Vec and BERT representations with supervised regressors; and full-conversation assessment with DeepSeek under zero-shot and few-shot prompting. The best LLM configuration achieved an accuracy of 0.851 and an F1-score of 0.788, while the audio approach showed comparatively balanced performance, and conventional text representations were less consistent. These findings illustrate the different information captured by acoustic, lexical, contextual, and generative approaches. Rather than advocating automated psychological diagnosis in educational settings, we use the results to discuss construct validity, domain transfer, interpretability, bias, privacy, and human oversight, arguing that socio-affective learning analytics should treat computational outputs as uncertain, contextual indicators for human interpretation.

Keywords: learning analytics; socio-affective analytics; speech analysis; large language models; affective computing

Copyright: © 202X by the authors. Submitted for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Nowadays, the digitalization of education has shifted learning environments into data-rich ecosystems where learning management systems, forums, intelligent tutoring systems, and AI-mediated tools continuously generate traces of learner activity. Learning analytics has traditionally relied on explicit behavioral indicators such as logins, clicks, time spent on platform, and submission patterns, which capture observable participation but only a limited dimension of the learner's experience.

Engagement, in particular, cannot be reduced to behavior. It involves behavioral, cognitive, emotional, and social dimensions, as well as persistence, self-regulation, and perceived agency (Bergdahl et al., 2024). Two students with similar interaction patterns may therefore be in very different learning conditions: frequent access does not imply cognitive engagement, and low activity does not imply disengagement. This mismatch between observable behavior and learner experience is a fundamental challenge for the field.

Recent AI advances make it possible to move beyond activity traces by exploiting written and spoken language, produced in forums, reflective diaries, presentations, tutoring dialogues, and interactions with conversational AI. These communicative acts may carry linguistic and paralinguistic signals of uncertainty, frustration, motivation, fatigue, or withdrawal, and multimodal learning analytics has begun to combine such heterogeneous sources (Ouhaichi et al., 2023). Voice conveys prosodic, temporal, and acoustic information (pitch, intensity, rate, pauses, spectral features) that can be modeled directly from spectrograms or raw audio content (Liu et al., 2024; Donaghy et al., 2024; Ji et al., 2024; Menne et al., 2024). Text can be represented at lexical, semantic, and contextual levels, from Word2Vec embeddings to BERT-style contextual models, enabling analysis of reduced elaboration, uncertainty markers, or affectively marked language (Chancellor & De Choudhury, 2020; Fisher et al., 2026).

Large Language Models (LLMs) add another layer: through prompts, they can summarize, classify, and interpret unstructured learner-generated content, for instance by surfacing recurrent difficulties or flagging passages for human attention (Shin et al., 2024; Omar & Levkovich, 2025; Dai et al., 2026). However, their interpretative flexibility does not imply reliability. Outputs are sensitive to prompting and configuration and may suffer from hallucination, hidden bias, and limited reproducibility. Their fluency may also lead users to attribute psychological understanding that the underlying mechanisms cannot support, which is especially problematic for socio-affective language (Bender et al., 2021).

The question is thus not only whether models detect statistically meaningful patterns, but which inferences institutions may legitimately and ethically draw from learner communication (Fareed et al., 2025). This matters when techniques from clinical contexts are transferred to education: a model that identifies acoustic or linguistic correlates of depressive symptoms in one dataset cannot simply become an instrument for labeling students as “depressed”, “fragile”, “at-risk”, or “disengaged”. Clinical and educational constructs, and the contexts in which they are observed, differ substantially (Alfredo et al., 2024; Khosravi et al., 2022; Prinsloo et al., 2023).

This paper critically reframes a comparative framework originally developed to automatically analyze depressive symptoms from voice and text, combining acoustic modeling, conventional and contextual textual representations, and LLM prompting. Rather than transferring its diagnostic aim, we use it as a methodological case to examine the opportunities and limits of voice- and language-based socio-affective learning analytics. The key distinction is that diagnostic analytics assigns individuals to clinical categories. In contrast, educational analytics should support situated interpretation and pedagogical decision-making, treating outputs as uncertain signals for human interpretation, not definitive assessments. This is consistent with closed-loop views of learning analytics, in which analysis must inform educational action

(Sailer et al., 2024); in the socio-affective case, closing the loop requires caution, since an output is meaningful only when interpreted in context, combined with other evidence, and translated into proportionate support.

The paper has three objectives: (1) to examine how acoustic, textual, and LLM-based approaches offer complementary representations of learner communication; (2) to discuss how methods developed for affective or psychological analysis can be reframed for education without becoming automated diagnosis; and (3) to identify methodological and ethical requirements for responsible use, namely construct validity, uncertainty, interpretability, privacy, bias, and human oversight.

Our broader argument is that the value of AI for socio-affective learning analytics depends not only on predictive performance but on the relation between computational inference and pedagogical interpretation. More and finer-grained data do not automatically yield deeper understanding of learning, and voice- and language-based analytics involve particularly sensitive forms of learner data and should therefore be designed around purpose limitation, transparency, human control, and meaningful stakeholder involvement (Liu & Khalil, 2023; Rets et al., 2023; Alfredo et al., 2024). Responsible learning analytics requires embedding technical models within an explicit educational and ethical framework.

2. Materials and Methods

This section describes the experimental framework adopted in the study, including the dataset and task definition, the speech- and text-based modeling pipelines, the LLM-based analysis, and the evaluation criteria used to compare the different approaches.

2.1. Dataset and Experimental Task

We conducted the experiments using the Distress Analysis Interview Corpus–Wizard of Oz (DAIC-WOZ)[1], which contains semi-structured human-computer interviews with audio recordings, textual transcripts, and depression-related clinical annotations (Gratch et al., 2014). The experimental partition comprised 142 training samples and 47 test samples, following the predefined dataset split.

We operationalized depression severity using the available PHQ-8 scores. For the common binary evaluation across the experimental branches, scores were binarized at a threshold of 10: participants with PHQ-8 scores of 10 or higher were assigned to the positive class, while lower scores were assigned to the negative class. Since the positive class was underrepresented, oversampling was applied to the training data to reduce class imbalance.

2.2. Speech-Based Analysis

The speech branch used a raw ECAPA-TDNN architecture incorporating Non-Uniform Speaker Disentanglement (NUSD)[2], following the approach proposed by Wang et al. (2023). Audio signals were cleaned to exclude irrelevant portions of the interviews, then normalized and standardized in length through

[1] https://dcapswoz.ict.usc.edu/

[2] https://github.com/kingformatty/NUSD/tree/main

EDIZIONI UNIVERSITARIE ROMANE
editoria scientifica dal 1952

padding or truncation. The training procedure included an auxiliary speaker-identification objective designed to encourage speaker-invariant representations. In addition to depression classification, the model therefore produced a speaker-related prediction and an intermediate embedding representation. Gradient disentanglement aims to reduce the contribution of speaker-specific characteristics to the representations used for depression classification (Ravi et al., 2024). Dynamic learning-rate adaptation and early stopping were applied during training.

2.3. Conventional Text-Based Analysis

Two alternative textual representations were investigated. We trained Word2Vec[3] embeddings using a skip-gram architecture with 300-dimensional vectors and a context window extending to five words around the target token. We obtained transcript-level representations by aggregating word embeddings into fixed-dimensional vectors. We obtained a second representation using a widespread pretrained model, i.e., BERT (Bidirectional Encoder Representations from Transformers). In detail, we adopted the *bert-base-uncased* model[4]. We tokenized the transcript content and processed it through BERT to obtain contextual sentence representations, which we then aggregated into fixed 768-dimensional vectors. The two representations therefore provided a comparison between classical distributional embeddings and richer contextual language representations. We used the resulting representations as input to Decision Tree, XGBoost, and AdaBoost regression models. We then converted continuous predictions into binary classifications using the PHQ-8 threshold described above.

2.4. LLM-Based Analysis

The third branch investigated whether an LLM could infer depression-related information directly from complete conversational transcripts without task-specific parameter training. We adopted the DeepSeek[5] model for each conversation and instructed it through a structured prompt to estimate the corresponding depression score. In particular, two prompting configurations were evaluated: (i) a zero-shot configuration, in which the task and scoring criteria were provided without labeled examples, and (ii) a few-shot configuration, in which the prompt additionally contained example input-output mappings intended to illustrate the scoring procedure. We explicitly distinguish these prompting strategies and use them to assess whether an LLM can support and improve the interpretation of clinically relevant linguistic cues.

2.5. Evaluation Metrics

We assessed performance using accuracy, precision, recall, and F1-score. Because the dataset is imbalanced and false positives and false negatives have different implications, we treated F1-score as a particularly relevant summary measure. For the audio branch, we reported both class-level and macro- and micro-aggregated metrics.

[3] https://code.google.com/archive/p/word2vec/

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

[5] https://api-docs.deepseek.com/

3. Results

This section reports the experimental results on the DAIC-WOZ dataset. Table 1 summarizes the performance obtained across the three analytical approaches considered in the study: audio-based analysis, embedding-based text analysis, and LLM prompting. The embedding-based configurations include Word2Vec and BERT representations combined with Decision Tree (DT), AdaBoost (ADA), and XGBoost (XGB), while the LLM-based analysis evaluates DeepSeek under zero-shot and few-shot prompting. For each configuration, we report accuracy, precision, recall, and F1-score to provide a comparative overview of predictive performance. The table therefore highlights the different behavior of acoustic, embedding-based, and generative approaches, providing the empirical basis for the comparative analysis discussed in the following sections.

Table 1. Classification results on the DAIC-WOZ dataset

Modality	Model	Accuracy	Precision	Recall	F1
Audio	ECAPA-TDNN	0.726	0.750	0.726	0.735
Embedding	BERT + DT	0.489	0.250	0.357	0.294
Embedding	BERT + XGB	0.702	0.500	0.357	0.418
Embedding	BERT + ADA	0.723	0.667	0.143	0.235
Embedding	Word2Vec + DT	0.617	0.400	0.571	0.471
Embedding	Word2Vec + XGB	0.723	1.000	0.071	0.133
Embedding	Word2Vec + ADA	0.383	0.308	0.857	0.453
LLM prompting	DeepSeek zero-shot	0.851	0.684	0.929	0.788
LLM prompting	DeepSeek few-shot	0.809	0.632	0.857	0.727

3.1. Speech-based Results

The speech-based model produced relatively balanced results compared with the conventional text pipelines. The macro-average F1-score of 0.735 indicates that raw acoustic information contained useful discriminative information, although performance remained lower for the depressed class than for the non-depressed class. This difference is particularly relevant given the original class imbalance and suggests that aggregate accuracy alone provides an incomplete account of model behavior.

3.2. Conventional Text-Based Results

Performance varied substantially across both representation methods and regressors. Specifically, for BERT representations, XGBoost achieved the highest F1-score, whereas AdaBoost achieved the highest accuracy but very low recall. With Word2Vec representations, Decision Tree provided the highest F1-score. XGBoost

reached perfect precision but a recall of only 0.071, indicating that it identified very few positive cases. These results indicate that, within this experimental configuration, the richer contextual representation offered by BERT did not automatically produce superior classification performance relative to Word2Vec.

3.3. LLM-based Results

The zero-shot DeepSeek configuration achieved the highest numerical performance among the evaluated configurations, while also identifying a large proportion of positive cases. Contrary to the initial expectation that examples might help interpret the scoring task, few-shot prompting did not improve performance. This finding suggests that the effectiveness of demonstrations depends not merely on their presence but on their representativeness, selection, and compatibility with the model's pre-existing representations.

4. Discussion

The comparative analysis shows that the three investigated strategies capture complementary dimensions of communicative data and exhibit markedly different predictive behaviors. Overall, the experiments highlight three main findings.

First, raw speech modeling achieved solid, comparatively balanced performance, supporting the relevance of paralinguistic information for socio-affective inference. Second, embedding-based text analysis produced more heterogeneous results, revealing a strong interaction among representation strategy, learning algorithm, and the precision-recall trade-off. Third, LLM prompting achieved the best overall performance among the investigated approaches. Taken together, these results support the value of analyzing communicative data at multiple representational levels, from acoustic signals to conventional semantic embeddings and discourse-level generative models.

The speech-based results reinforce the established role of speech signals in affective computing, showing that useful information for affective inference is not restricted to the lexical content of an interaction. Acoustic and paralinguistic characteristics can provide substantial discriminative information even in the absence of textual information. From a learning analytics perspective, this result is particularly relevant in educational interactions that frequently involve oral communication through presentations, tutoring sessions, collaborative activities, oral examinations, spoken reflections, and increasingly voice-based interfaces.

Concerning embedding-based experiments, they provide another relevant insight. Rather than showing a uniform contribution of textual information, the results reveal substantial differences in how individual combinations of representations and learning algorithms behave. These results show that the effectiveness of textual representations cannot be assessed independently of the most employed learning strategy and the evaluation objective. A particularly interesting finding is that BERT did not consistently outperform Word2Vec despite providing richer contextual representations. This should not necessarily be interpreted as evidence against contextual embeddings. Instead, it shows that greater representational complexity does not automatically translate into improved performance under every experimental condition. Dataset size, feature aggregation,

EDIZIONI UNIVERSITARIE ROMANE
editoria scientifica dal 1952

classifier characteristics, and task formulation can all affect the benefit obtained from richer representations. Methodologically, this finding strengthens the value of comparative experimentation rather than assuming that more recent representation techniques are intrinsically superior.

The strongest finding concerns the LLM-based prompting approach, which yielded the best overall results. In particular, DeepSeek in the zero-shot setting achieved the best performance, with, in contrast with most compared models, a remarkably high recall, indicating a strong capacity to identify positive cases in the experimental dataset. The few-shot configuration also performed competitively, although it did not improve on the zero-shot setting. This is important because the LLM was not fine-tuned on the experimental dataset; instead, it was prompted to interpret complete conversational transcripts. The results therefore suggest that large-scale language pre-training can provide representations that exploit complex semantic, pragmatic, and discourse-level information that may be difficult to capture through fixed transcript embeddings.

The comparison between zero-shot and few-shot prompting is itself informative. The experiments did not support the expectation that demonstrations would necessarily facilitate the task. Instead, the zero-shot configuration performed better across the principal metrics. One plausible interpretation is that the pretrained model already encoded sufficiently rich linguistic knowledge to address the task effectively, while the selected few-shot examples may have constrained or biased the model toward a narrower set of linguistic patterns. More generally, this result indicates that the design and selection of demonstrations should be treated as an experimental variable rather than as an automatically beneficial component of LLM prompting.

Taken together, these findings suggest a progression in how communicative information is represented. The audio-based approach captures acoustic and paralinguistic information; embedding-based methods represent lexical and contextual semantic information through fixed numerical representations; and LLM prompting operates on complete conversational contexts, potentially capturing relationships among statements, semantic nuances, and discourse-level patterns. The higher performance of the LLM approach therefore does not diminish the relevance of the other strategies. On the contrary, the results suggest that these approaches offer complementary perspectives on socio-affective information and motivate future development of architectures that integrate, rather than merely compare, these sources.

At the same time, transferring these methods from depression-related analysis to education requires a change in the target of inference rather than a simple reuse of the original classifier. Clinical depression and educational engagement are distinct constructs, and the present experiments do not claim to validate depression-detection models as instruments for assessing students. Instead, this work demonstrates the computational potential of different communicative representations and uses this evidence to identify promising methodological directions for socio-affective learning analytics.

In this perspective, the distinction between diagnostic and educational analytics becomes crucial. Rather than assigning learners definitive psychological labels, future systems could use voice- and language-derived signals to identify patterns relevant to participation, confidence, uncertainty, frustration, persistence, or changes

in engagement. Such signals could complement existing learning analytics indicators and support educators in identifying situations that may warrant further attention. This interpretation aligns with the framework's original educational reframing, in which computational outputs serve as inputs to situated human interpretation rather than autonomous diagnoses.

The results also have implications for constructing ground truth in future educational studies, in which validation may combine multiple evidence sources, including learner self-reports, behavioral traces, teacher or tutor observations, qualitative assessments, and linguistic or acoustic indicators.

The strong performance of LLM prompting also opens opportunities beyond classification. Because LLMs operate directly on natural language, they could potentially support qualitative forms of analytics, for example by summarizing longitudinal changes in learner discourse, highlighting recurring expressions of uncertainty, or organizing large quantities of reflective text for teacher inspection.

Finally, the results support a human-in-the-loop perspective in which computational approaches extend educators' analytical capacity rather than replace human judgment. The complementary behavior observed across audio, embedding-based, and LLM approaches suggests that socio-affective analytics may be most valuable when it provides multiple forms of evidence that can be interpreted within the broader educational context.

Overall, the study shows that communicative data contains exploitable information at multiple levels of representation and that recent LLM-based approaches can substantially extend the capabilities of conventional acoustic and textual analysis. The findings provide both empirical evidence for the effectiveness of the investigated methods and a foundation for their responsible extension toward socio-affective learning analytics. Future work should therefore focus on validating these approaches on educational data, integrating complementary modalities, and investigating how to incorporate predictive and generative models into transparent, human-centered analytical workflows.

5. Conclusions

This study compared speech-based, conventional text-based, and Large Language Model approaches within a challenging socio-affective inference task. The results showed that raw speech modeling retained useful discriminative information, conventional textual representations produced heterogeneous and generally weaker results, and an LLM analyzing complete conversations achieved the highest reported F1-score, with zero-shot prompting outperforming the tested few-shot configuration.

Beyond their original clinical setting, these results provide a useful methodological case for considering the role of voice and language in learning analytics. They show that different computational representations capture different dimensions of human communication, while also exposing the limits of treating such representations as direct measures of latent affective states.

The educational relevance of these technologies should therefore not be understood as an opportunity to automate psychological assessment of students. Instead, future socio-affective learning analytics systems should use computational

models to provide uncertain and contextual evidence that complements, rather than replaces, human interpretation. Validation on educational data, multimodal triangulation, transparency, privacy protection, bias assessment, and meaningful human oversight are essential prerequisites for such a transition.

References

Alfredo, R., Echeverria, V., Jin, Y., Yan, L., Swiecki, Z., Gašević, D., & Martinez-Maldonado, R. (2024). *Human-centred learning analytics and AI in education: A systematic literature review*. Computers and Education: Artificial Intelligence, 6, 100215. DOI: 10.1016/j.caeai.2024.100215.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). *On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?* FAccT '21, 610–623. DOI: 10.1145/3442188.3445922.

Bergdahl, N., Bond, M., Sjöberg, J., Dougherty, M., & Oxley, E. (2024). *Unpacking student engagement in higher education learning analytics: A systematic review*. International Journal of Educational Technology in Higher Education, 21, 63. DOI: 10.1186/s41239-024-00493-y.

Chancellor, S., & De Choudhury, M. (2020). *Methods in predictive techniques for mental health status on social media: A critical review*. npj Digital Medicine, 3, 43. DOI: 10.1038/s41746-020-0233-7.

Dai, A., Shi, W., Gu, X., & Huang, L. (2026). *Large Language Models for Depression Detection: A Review with Prospects of Incomplete Multimodality*. Brain Sciences, 16(6), 593. DOI: 10.3390/brainsci16060593.

Donaghy, P., Ennis, E., Mulvenna, M., et al. (2024). *A Review of Studies Using Machine Learning to Detect Voice Biomarkers for Depression*. Journal of Technology in Behavioral Science. DOI: 10.1007/s41347-024-00454-2.

Fareed, M., Fatima, M., Uddin, J., Ahmed, A., & Sattar, M. A. (2025). *A systematic review of ethical considerations of large language models in healthcare and medicine*. Frontiers in Digital Health, 7, 1653631. DOI: 10.3389/fdgth.2025.1653631.

Fisher, H., Jaffe, N. M., Pidvirny, K., et al. (2026). *Language-based detection of depression with machine learning: systematic review and meta-analysis*. npj Digital Medicine, 9, 273. DOI: 10.1038/s41746-026-02448-1.

Gratch J, Artstein R, Lucas GM, Stratou G, Scherer S, Nazarian A, Wood R, Boberg J, DeVault D, Marsella S, Traum DR. *The Distress Analysis Interview Corpus of Human and Computer Interviews*. In Proceedings of LREC 2014 May (pp. 3123-3128).

Ji, J., Dong, W., Li, J., et al. (2024). *Depressive and mania mood state detection through voice as a biomarker using machine learning*. Frontiers in Neurology, 15, 1394210. DOI: 10.3389/fneur.2024.1394210.

Khosravi, H., Buckingham Shum, S., Chen, G., et al. (2022). *Explainable Artificial Intelligence in education*. Computers and Education: Artificial Intelligence, 3, 100074. DOI: 10.1016/j.caeai.2022.100074.

Liu, L., Liu, L., Wafa, H. A., Tydeman, F., Xie, W., & Wang, Y. (2024). *Diagnostic accuracy of deep learning using speech samples in depression: A systematic review and meta-analysis*. JAMIA, 31(10), 2394–2404. DOI: 10.1093/jamia/ocae189.

Liu, Q., & Khalil, M. (2023). *Understanding privacy and data protection issues in learning analytics using a systematic review*. British Journal of Educational Technology, 54, 1715–1747. DOI: 10.1111/bjet.13388.

Menne, F., Dörr, F., Schräder, J., et al. (2024). *The voice of depression: speech features as biomarkers for major depressive disorder*. BMC Psychiatry, 24, 794. DOI: 10.1186/s12888-024-06253-6.

Omar, M., & Levkovich, I. (2025). *Exploring the efficacy and potential of large language models for depression: A systematic review*. Journal of Affective Disorders, 371, 234–244. DOI: 10.1016/j.jad.2024.11.052.

Ouhaichi, H., Spikol, D., & Vogel, B. (2023). *Research trends in multimodal learning analytics: A systematic mapping study*. Computers and Education: Artificial Intelligence, 4, 100136. DOI: 10.1016/j.caeai.2023.100136.

Prinsloo, P., Slade, S., & Khalil, M. (2023*). Multimodal learning analytics - In-between student privacy and encroachment: A systematic review*. British Journal of Educational Technology, 54(6). DOI: 10.1111/bjet.13373.

Ravi, V., Wang, J., Flint, J., & Alwan, A. (2024). *Enhancing accuracy and privacy in speech-based depression detection through speaker disentanglement*. Computer Speech & Language, 86, 101605. DOI: 10.1016/j.csl.2023.101605.

Rets, I., Herodotou, C., & Gillespie, A. (2023). *Six Practical Recommendations Enabling Ethical Use of Predictive Learning Analytics in Distance Education*. Journal of Learning Analytics, 10(1), 149–167. DOI: 10.18608/jla.2023.7743.

Sailer, M., Ninaus, M., Huber, S. E., Bauer, E., & Greiff, S. (2024). *The End is the Beginning is the End: The closed-loop learning analytics framework*. Computers in Human Behavior, 158, 108305. DOI: 10.1016/j.chb.2024.108305.

Shin, D., Kim, H., Lee, S., Cho, Y., & Jung, W. (2024*). Using Large Language Models to Detect Depression From User-Generated Diary Text Data as a Novel Approach in Digital Mental Health Screening*. Journal of Medical Internet Research, 26, e54617. DOI: 10.2196/54617.

Wang, J., Ravi, V., & Alwan, A. (2023). *Non-uniform Speaker Disentanglement For Depression Detection From Raw Speech Signals*. Interspeech 2023, 2343–2347. DOI: 10.21437/Interspeech.2023-2101.