Digital biomarkers of mood states from speech in bipolar disorder

Cristina Crocamo[1,*], Aurelia Canestro[1], Dario Palpella[1], Riccardo M. Cioni[1], Christian Nasti[1], Susanna Piacenti[1], Alessandra Bartoccetti[1], Martina Re[1], Valentina Simonetti[2], Chiara Barattieri di San Pietro[2], Maria Bulgheroni[2], Francesco Bartoli[1] and Giuseppe Carrà[1]

[1]*School of Medicine and Surgery, University of Milano-Bicocca, Monza, Italy.*
[2]*Ab.Acus s.r.l. Milan, Italy.*

Abstract

Regular monitoring is essential to effectively track mood fluctuations and assess ongoing treatment needs for mood disorders (e.g., identifying early signs of relapse, adjusting therapeutic interventions, and improving long-term outcomes). The current ongoing work aims at assessing the relationships between language and symptom severity in people with bipolar disorders, thus investigating potential mHealth mood detection mechanisms based on speech patterns. Acoustic features included conversational measures for nonverbal language and statistics for prosodic cues. Preliminary results, combining acoustic features and natural language processing (NLP) scores, were promising, somehow discriminating clinical conditions of people with BD when assessing their mood states. This approach may offer potential benefits for individualized mental health care and early intervention approaches in real-world scenarios.

Keywords

speech, signal analysis, mood states, mHealth, remote assessment, machine learning, neural network

1. Introduction

Bipolar disorder (BD) is a lifelong episodic illness resulting in reduced psychosocial functioning. The majority of BD cases onset in early adulthood and it is among the leading causes of disability in working-age adults [1]. Community services often struggle in delivering regular monitoring of treatment needs, contributing to a gap in care [2]. The assessment of mood states and potential variations is pivotal in BD [3, 4]. Because of its chronicity, approaches for prediction and prevention of further episodes in which the patient's mood and activity levels are considerably disturbed are critical [4]. Mood states are defined referring to the presence and severity of depressive symptoms as assessed by the Montgomery-Åsberg Depression Rating Scale (MADRS), including items that measure sadness feelings [5], and manic symptoms as assessed by the Young Mania Rating Scale (YMRS), measuring elevated mood and increased activity levels [6]. Higher scores indicate more severe depressive or manic symptoms, respectively.

Traditionally, this evaluation has heavily relied on clinical interviews, including an analysis of thoughts and their manifestation in language of people with BD. Focusing on meaning and communication, language has a central role for diagnosis and treatment in BD with speech patterns being

Italian Workshop on Artificial Intelligence for Human Machine Interaction (AIxHMI 2024), November 26, 2024, Bolzano, Italy
*Corresponding author.
cristina.crocamo@unimib.it (C. Crocamo); a.canestro@campus.unimib.it (A. Canestro); d.palpella@campus.unimib.it (D. Palpella); r.cioni1@campus.unimib.it (R. M. Cioni); c.nasti@campus.unimib.it (C. Nasti); s.piacenti1@campus.unimib.it (S. Piacenti); a.bartoccetti@campus.unimib.it (A. Bartoccetti); m.re22@campus.unimib.it (M. Re); valentinasimonetti@ab-acus.eu (V. Simonetti); cbarattieri@fatebenefratelli.eu (C. B. d. S. Pietro); mariabulgheroni@ab-acus.com (M. Bulgheroni); francesco.bartoli@unimib.it (F. Bartoli); giuseppe.carra@unimib.it (G. Carrà)
https://en.unimib.it/cristina-crocamo (C. Crocamo); https://en.unimib.it/francesco-bartoli (F. Bartoli); https://en.unimib.it/giuseppe-carra (G. Carrà)
0000-0002-2979-2107 (C. Crocamo); 0000-0003-2718-4076 (A. Canestro); 0000-0002-3073-8407 (D. Palpella); 0000-0001-5718-9930 (R. M. Cioni); 0000-0003-2986-2125 (C. Nasti); 0000-0002-8184-2634 (S. Piacenti); 0000-0002-0862-8703 (A. Bartoccetti); 0000-0003-4385-2951 (M. Re); 0000-0001-7013-1338 (V. Simonetti); 0000-0003-4407-7037 (C. B. d. S. Pietro); 0000-0001-8484-9834 (M. Bulgheroni); 0000-0003-2612-4119 (F. Bartoli); 0000-0002-6877-6169 (G. Carrà)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

crucial when assessing current experiences, emotions, and thoughts. For instance, pressure of speech encompassing a high number of words during phonation and a small number of pauses is likely to be a sign of underlying manic symptoms. Conversely, mood states in depression are characterized by poverty of speech and a monotone pitch [7, 8, 9, 10]. Therefore, we hypothesized that speech patterns would relate to standard psychometric assessments in BD, discriminating clinical conditions when predicting individual mood states.

2. Related work

Progress in Machine Learning (ML) and Natural Language Processing (NLP) techniques may support the development of automated systems assessing speech patterns as objective markers of mood states [11]. A recent review highlighted favourable evidence about the use of audio data to monitor mood disorders, despite some challenges [12]. However, prior research emphasized the potential for the use of speech mainly to distinguish between individuals with and without a variety of psychiatric disorders, including BD [9, 13]. Alternatively, some studies focused on the correlation between acoustic features and parameters from electroglottographic signal of voiced segments [14]. Acoustic features (e.g., jitter) were identified as likely reflecting a dysregulation of autonomous nervous system that influence muscular tone and articulatory control. Consistently, specifically considering mood fluctuations among people with BD, available evidence suggests that speech patterns impairments may be sensitive and valid measures of mood states [15, 16]. Previous work on ecological speech signal analysis from phone calls recordings showed model ability to differentiate between hypomanic and euthymic as well as between depressed and euthymic speech according to a support vector machine classifier (average AUC of 0.81 for hypomania and 0.67 for depression) [17]. Similarly, a more recent study based on phone calls data trained ML models considering random forest classifiers to classify mood states according to estimated voice features [16]. Although a varying accuracy (0.61 to 0.74) was estimated when classifying a depressive or manic versus a euthymic state, these approaches seem promising to complement rating scales with speech markers, thus possibly improving mood states monitoring in real-world settings.

However, there are several barriers in the implementation of mood detection systems in real-world applications, including high degree of heterogeneity between studies and the use of non-standardized metrics reporting [11]. Moreover, several areas remain understudied, including the use of speech spectrograms testing performance to remotely assess the individual clinical status among people with BD. However, a few studies explored speech-based emotion recognition using spectrograms in related fields. A recent study proposed a convolutional neural network for anger and stress detection using handcrafted features and deep learned features [18]. A further study explored speech emotion recognition from the utterances of interacting professional actors performing spontaneously, by exploiting a novel convolutional neural network architecture to recognize speech emotions based on local correlations and global contextual information from speech spectrograms [19]. These approaches have been proven successful, with high accuracy for speech emotion recognition, emphasizing related feasibility when processing speech segments. In addition, these systems can be integrated with other signals (i.e., linguistic and paralinguistic components of speech), thus implementing the analysis of a multimodal signal [19].

3. Speech signal analysis and deep learning for mood prediction

Referring to the process of examining and interpreting the characteristics of spoken output, speech signal analysis was used to identify key patterns from acoustic signals generated during speech. Speech signal analysis was proven effective to characterize mood states of people with BD, thus contributing to individualized approaches including the estimate of various speech features based on signal's frequency and energy/amplitude [20]. Remarkably, deep learning techniques may significantly advance speech signal processing, enabling more accurate recognition, analysis, and interpretation of individuals' language, especially for mood detection. Indeed, deep learning encompasses ML techniques that can automatically learn hierarchical representations from data. Core models embrace neural networks (NN) including multiple interconnected layers of nodes (mimicking the structure and functioning of the human brain) as well as convolutional neural networks (CNN), in which the nodes of each layer are clustered [21]. In addition, considering spectrograms in speech analysis systems, the latter may foster automatically learning representations from image data according to a benchmark performance in image classification [19].

3.1. Proposed approach

Assessing mood fluctuations based on gold-standard assessments of mania and depression in BD, the proposed approach aimed at exploiting deep learning algorithms for mood states prediction. Eligible participants involved subjects with a diagnosis of BD, aged between 18 and 65 years old from both inpatient and outpatient services. They were approached by specifically trained staff. Subjects unable to provide informed consent, those with vocal or hearing issues were excluded. Speech data were collected and processed through a mobile app that study participants accessed on their smartphones using password-protected access to self-administer verbal performance tasks. In particular, the system embedded in the smartphone was grounded on a cloud-based architecture hosting the system database, the Representational State Transfer (REST) Application Programming Interfaces (API), and the backend processing modules.

Considering mood states variability and relevant speech signal segmentation, we aimed to combine two different approaches for speech analysis. First, highlighting speaking segments from raw audio data, speech was automatically processed through speech recognition and quantities representing voice characteristics (i.e., acoustic features) were estimated. By leveraging Parselmouth module as a bridge to speech-to-text preprocessing and related Praat's built-in functions, basic acoustic features were computed (e.g., fundamental frequency, harmonics-to-noise ratio, jitter and shimmer). Speech rate, verbal task duration, and phonation duration were also considered. In addition, acoustic features were integrated with both standard and novel NLP scores for linguistic components of speech according to distributional semantic models (e.g., estimating information on processing speed and capturing both lexical overlap and semantic similarity in the spoken output). As a whole, previous studies showed an enhanced performance considering combined features [22]. With predictive accuracy as the primary goal rather than understanding the exact contributions of individual features (both speech signal-derived and NLP-extracted features), the models' architecture was based on a feedforward neural network with fully connected layers (Rectified Linear Unit -ReLU- activation in two hidden layers and sigmoid activation in the output layer). The Adaptive Moment Estimation (adam) optimizer was used with the model seeking to minimize cross-entropy loss. The model stops the training if the validation loss does not improve for 10 epochs. A 5-fold cross-validation was used with each fold providing metrics that are averaged for the final evaluation based on 80% of the data for training and 20% for testing.

On the other hand, by considering the feasibility of performing acoustic signalling analysis as a function of frequency and time, this ongoing work aimed to focus on potential speech segments from spectrograms for image data classification. Higher energy against low energy regions (e.g., pauses) in spectrograms may be distinguished by darker/lighter colours. Consistently, spectrograms may be able to display the properties of a changing signal through a series of snapshots according to segment length with speech corpora capturing tones, emotions, rhythms among signals beyond content of speech.

4. Results

Based on the caseload of the ASST Nord Milano Mental Health Care Trust, 37 subjects with BD were involved, while enrolment is still active. Involved participants mainly lived alone or with family, were globally educated, but unemployed. They were likely to report severe depressive symptoms (MADRS score ≥19, 46%) and just a few had severe manic features (YMRS score ≥20; 24%). Therefore, we preliminarily focused on MADRS assessment (Table 1), taking into account sex-specific discriminating ability of NLP-based and acoustic features for mood states prediction.

Table 1
MADRS items.

Item	Rating description
Apparent Sadness	Representing despondency, gloom and despair, (more than just ordinary transient low spirits) reflected in speech, facial expression, and posture. Rated by depth and inability to brighten up.
Reported Sadness	Representing reports of depressed mood, regardless of whether it is reflected in appearance or not. Includes low spirits, despondency or the feeling of being beyond help and without hope. Rated according to intensity, duration and the extent to which the mood is reported to be influenced by events.
Inner Tension	Representing feelings of ill-defined discomfort, edginess, inner turmoil, mental tension mounting to either panic, dread or anguish. Rated according to intensity, frequency, duration and the extent of reassurance called for.
Reduced Sleep	Representing the experience of reduced duration or depth of sleep compared to the subject's own normal pattern when well.
Reduced Appetite	Representing the feeling of a loss of appetite compared with when well. Rated by loss of desire for food or the need to force oneself to eat.
Concentration Difficulties	Representing difficulties in collecting one's thoughts mounting to incapacitating lack of concentration. Rated according to intensity, frequency, and degree of incapacity produced.
Lassitude	Representing a difficulty getting started or slowness initiating and performing everyday activities.
Inability to Feel	Representing the subjective experience of reduced interest in the surroundings, or activities that normally give pleasure. The ability to react with adequate emotion to circumstances or people is reduced.
Pessimistic Thoughts	Representing thoughts of guilt, inferiority, reproach, sinfulness, remorse and ruin.
Suicidal Thoughts	Representing the feeling that life is not worth living, that a natural death would be welcome, suicidal thoughts, and preparations for suicide. Suicidal attempts should not in themselves influence the rating.

Table 2
Results for mood prediction in BD based on neural network models, stratified by sex.

Selected features	MADRS					
	Overall		**Female**		**Male**	
	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC
NLP-based	0.707	0.850	0.750	0.800	0.767	0.700
Acoustic (F0, jitter- and shimmer-related only)	0.543	0.613	0.500	0.667	0.500	0.750
Acoustic including speech rate, duration, phonation duration	0.650	0.703	0.750	0.683	0.800	0.650
Combined	0.711	0.647	0.750	0.750	0.467	0.650

Model performance was comprehensively evaluated according to accuracy and Receiver Operating Characteristic (ROC) Area Under the Curve (AUC) estimates to assess the ability of the models to correctly classify mood states. In particular, two classes for symptom severity were considered for classification (i.e., severe/not severe). Neural network models developed -including different sets of speech features and considering a chance level of 0.5 - showed varying levels of performance (Table 2). Notably, NLP features, such as mean intraword time and semantic similarity between words, provided satisfactory results as compared to models relying on acoustic features only (e.g., fundamental frequency, and jitter- and shimmer-related features). However, further analysis revealed that sex-based differences influenced the models' ability to accurately discriminate between mood states, thus suggesting that sex may modulate the expression of mood in both linguistic and acoustic features.

Figure 1 shows sample speech spectrograms of two study participants with different levels of symptom

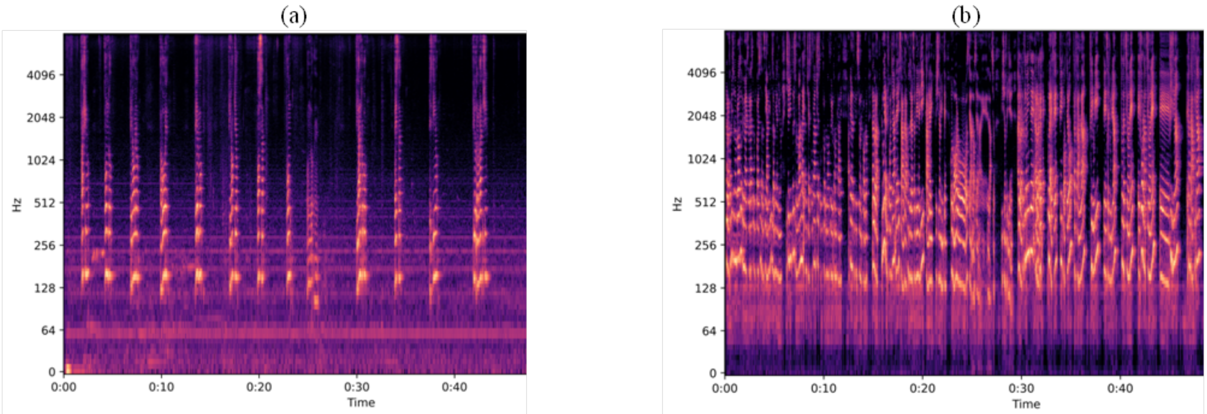

Figure 1: Speech spectrograms of people with BD; a) scoring high at MADRS (MADRS score ≥19) and low at YMRS; b) scoring low at MADRS and high at YMRS (YMRS score ≥20).

severity. While sampled spectrograms might have relatively limited representativeness, related visual quality -based on the identification of key features- can indicate how well this captures the underlying patterns in the data, thus providing a useful representation of the signal for further analyses. Indeed, visual inspection of relevant spectrograms of people with BD revealed likely distinct acoustic patterns when assessing mood states, possibly reflecting symptom severity. Specifically, sample patterns from verbal tasks of participants were likely to exhibit a different number and duration of pauses with varying speech rate and mean intraword time (i.e., presence/absence of pressure of speech) as well as potential differences in signal frequency and intensity. However, according to existing evidence, no standardized feature framework is available. Therefore, considering the uncertainty about which features should be extracted as well as the risk of bias due to potentially missing information, these results suggested the need to focus on speech segments more in detail, by pre-processing and analyse image data directly from speech spectrograms for image data classification purposes, possibly corroborating the role of speech features as digital markers of mood states in people with BD.

5. Conclusions and future research

The current work explored the use of speech signal analysis to map symptom severity in people with BD when assessing mood states using neural networks. Preliminary results showed relatively adequate accuracy for prediction, though with varying model performance according both to features selected and subgroups (e.g., sex). As a whole, combining acoustic signals and NLP can be a feasible, clinically useful, application in mental healthcare with acoustic features representing novel markers for mood states.

Future work will focus on capturing a higher degree of complexity of the underlying data distribution, by extending in a larger, more diverse sample of people with BD, accounting for potential confounders, and exploring speech segments more in detail based on speech spectrograms. This would enable a better understanding of how these methods can operate in real-world settings, particularly with regard to their potential for integration into clinical practice. Indeed, feature information may be complemented by feeding the spectrograms directly into the models as input data, using short voice segments to develop deep learning algorithms from speech spectrograms for classification purposes (e.g., CNN). In addition, speech signals are often mixed with other signals and both frequency and amplitude are likely to change over time resulting in non-stationary and non-linear signals. Therefore, empirical mode decomposition (EMD) related approaches, by breaking the signal down into components that reflect relevant changes, could offer valuable insights [23, 24, 25]. Consistently, smartphone-based approaches for speech processing show potential for real-time monitoring (or detection) of mood states in BD likely relying on ecological momentary assessments with NLP and artificial intelligence (AI) being promising for smart mental healthcare over time [26, 27, 28]. Based on mHealth technologies, this approach would help devising human-centered mood remote monitoring based on symptom patterns from speech possibly with significant clinical impact.

References

[1] S. Bolton, J. Warner, E. Harriss, J. Geddes, K. E. A. Saunders, Bipolar disorder: Trimodal age-at-onset distribution, Bipolar Disord 23 (2021) 341–356. doi:10.1111/bdi.13016.

[2] M. Bauer, O. A. Andreassen, J. R. Geddes, L. V. Kessing, U. Lewitzka, T. G. Schulze, E. Vieta, Areas of uncertainties and unmet needs in bipolar disorders: clinical and research perspectives, Lancet Psychiatry 5 (2018) 930–939. doi:10.1016/S2215-0366(18)30253-0.

[3] I. Grande, M. Berk, B. Birmaher, E. Vieta, Bipolar disorder, Lancet 387 (2016) 1561–1572. doi:10.1016/S0140-6736(15)00241-X.

[4] R. S. McIntyre, M. Berk, E. Brietzke, B. I. Goldstein, C. López-Jaramillo, L. V. Kessing, G. S. Malhi, A. A. Nierenberg, J. D. Rosenblat, A. Majeed, E. Vieta, M. Vinberg, A. H. Young, R. B. Mansur, Bipolar disorders, Lancet 396 (2020) 1841–1856. doi:10.1016/S0140-6736(20)31544-0.

[5] S. A. Montgomery, M. Asberg, A new depression scale designed to be sensitive to change, Br J Psychiatry 134 (1979) 382–389. doi:10.1192/bjp.134.4.382.

[6] R. C. Young, J. T. Biggs, V. E. Ziegler, D. A. Meyer, A rating scale for mania: reliability, validity and sensitivity, Br J Psychiatry 133 (1978) 429–435. doi:10.1192/bjp.133.5.429.

[7] D. Harvey, F. Lobban, P. Rayson, A. Warner, S. Jones, Natural language processing methods and bipolar disorder: Scoping review, JMIR Ment Health 9 (2022) e35928. doi:10.2196/35928.

[8] A. C. Arevian, D. Bone, N. Malandrakis, V. R. Martinez, K. B. Wells, D. J. Miklowitz, S. Narayanan, Clinical state tracking in serious mental illness through computational analysis of speech, PLoS One 15 (2020) e0225695. doi:10.1371/journal.pone.0225695.

[9] M. Faurholt-Jepsen, D. A. Rohani, J. Busk, M. L. Tønning, M. Vinberg, J. E. Bardram, L. V. Kessing, Discriminating between patients with unipolar disorder, bipolar disorder, and healthy control individuals based on voice features collected from naturalistic smartphone calls, Acta Psychiatr Scand 145 (2022) 255–267. doi:10.1111/acps.13391.

[10] J. C. Mundt, A. P. Vogel, D. E. Feltner, W. R. Lenderking, Vocal acoustic biomarkers of depression severity and treatment response, Biol Psychiatry 72 (2012) 580–587. doi:10.1016/j.biopsych.2012.03.015.

[11] N. Cummins, A. Baird, B. W. Schuller, Speech analysis for health: Current state-of-the-art and the increasing impact of deep learning, Methods 151 (2018) 41–54. doi:10.1016/j.ymeth.2018.07.007.

[12] F. Or, J. Torous, J. P. Onnela, High potential but limited evidence: Using voice data from smartphones to monitor and diagnose mood disorders, Psychiatr Rehabil J 40 (2017) 320–324. doi:10.1037/prj0000279.

[13] J. N. de Boer, A. E. Voppel, M. J. H. Begemann, H. G. Schnack, F. Wijnen, I. E. C. Sommer, Clinical use of semantic space models in psychiatry and neurology: A systematic review and meta-analysis, Neurosci Biobehav Rev 93 (2018) 85–92. doi:10.1016/j.neubiorev.2018.06.008.

[14] N. Vanello, A. Guidi, C. Gentili, S. Werner, G. Bertschy, G. Valenza, A. Lanata, E. P. Scilingo, Speech analysis for mood state characterization in bipolar patients, in: Annu Int Conf IEEE Eng Med Biol Soc, 2012, pp. 2104–2107. doi:10.1109/EMBC.2012.6346375.

[15] K. Matton, M. G. McInnis, E. M. Provost, Into the wild: Transitioning from recognizing mood in clinical interactions to personal conversations for individuals with bipolar disorder, in: Interspeech, Graz, Austria, 2019. doi:10.21437/Interspeech.2019-2698.

[16] M. Faurholt-Jepsen, J. Busk, M. Frost, M. Vinberg, E. M. Christensen, O. Winther, J. E. Bardram, L. V. Kessing, Voice analysis as an objective state marker in bipolar disorder, Transl Psychiatry 6 (2016). doi:10.1038/tp.2016.123.

[17] Z. N. Karam, E. M. Provost, S. Singh, J. Montgomery, C. Archer, G. Harrington, M. G. Mcinnis,

Ecologically valid long-term mood monitoring of individuals with bipolar disorder using speech, in: Proc. IEEE Int. Conf. Acoust. Speech Signal Process, 2014, pp. 4858–4862. doi:10.1109/ICASSP.2014.6854525.

[18] S. Kapoor, T. Kumar, Fusing traditionally extracted features with deep learned features from the speech spectrogram for anger and stress detection using convolution neural network, Multimed Tools Appl 81 (2022) 31107–31128. doi:10.1007/s11042-022-12886-0.

[19] H. Meng, T. Yan, F. Yuan, H. Wei, Speech emotion recognition from 3D log-mel spectrograms with deep learning network, IEEE Access 7 (2019) 125868–125881. doi:10.1109/ACCESS.2019.2938007.

[20] A. Z. Antosik-Wójcińska, M. Dominiak, M. Chojnacka, K. Kaczmarek-Majer, K. R. Opara, W. Radziszewska, A. Olwert, Ł. Święcicki, Smartphone as a monitoring tool for bipolar disorder: a systematic review including data analysis, machine learning algorithms and predictive modelling, Int J Med Inform 138 (2020). doi:10.1016/j.ijmedinf.2020.104131.

[21] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (2015) 436–444. doi:10.1038/nature14539.

[22] H. Naderi, B. H. Soleimani, S. Matwin, Multimodal deep learning for mental disorders prediction from audio speech samples, in: 33rd Conference on Neural Information Processing Systems (NeurIPS), Vancouver, Canada, 2019, pp. 4858–4862. arXiv 1909.01067v5.

[23] P. Marti-Puig, E. Gallego-Jutglà, G. Masferrer, J. Solé-Casals, A New Algorithm for Speech Enhancement Based on Multivariate Empirical Mode Decomposition, Artificial Intelligence Research and Development, IOS Press., 2018, pp. 247–255. doi:10.3233/978-1-61499-918-8-247.

[24] C. Sun, H. Li, L. Ma, Speech emotion recognition based on improved masking EMD and convolutional recurrent neural network, Front Psychol 13 (2023) 1075624. doi:10.3389/fpsyg.2022.1075624.

[25] U. Souza, J. Escola, T. Vedovatto, L. Brito, R. Lemos, Bidirectional EMD-RLS: Performance analysis for denoising in speech signal, Journal of Computational Science 74 (2023) 102181. doi:10.1016/j.jocs.2023.102181.

[26] E. Kerz, S. Zanwar, Y. Qiao, D. Wiechmann, Toward explainable AI (XAI) for mental health detection based on language behavior, Front Psychiatry 14 (2023) 1219479. doi:10.3389/fpsyt.2023.1219479.

[27] B. Zhou, G. Yang, Z. Shi, S. Ma, Natural language processing for smart healthcare, IEEE Rev Biomed Eng 17 (2024) 4–18. doi:10.1109/RBME.2022.3210270.

[28] W. Hinzen, L. Palaniyappan, The 'L-factor': Language as a transdiagnostic dimension in psychopathology, Prog Neuropsychopharmacol Biol Psychiatry 131 (2024) 110952. doi:10.1016/j.pnpbp.2024.110952.