Sanora: A Conversational AI Agent for Multimodal Digital Biomarkers of Mental Health

Matthew So[1,*], Michael Sobolev[2] and Gregory Menvielle[1]

[1] Okaya
[2] Cornell Tech

Abstract

Multimodal data (video, audio, text) can yield digital biomarkers relevant to depression and overall mental health. This study aimed to (1) extract an evidence-based library of vision, speech, and language biomarkers; (2) assess the feasibility of a fully remote conversational platform (Okaya) for collecting multimodal data; and (3) conduct preliminary signal checks for depression, fatigue, and cognition. Participants (N=10) were recruited from a women's mental health group in Australia and completed a total of 59 sessions. During the study, participants completed a check-in via Sanora, a conversational AI agent with an integration to a large-language model (LLM). From these interactions, 66 visual, acoustic, and text features were extracted. Validated assessments were also collected: PHQ-9 (depression), Cancer Fatigue Scale (fatigue), and Trail Making Test (cognition). We evaluated correlations between extracted features and assessments. Preliminary correlations identified promising digital biomarkers for depression (average F5 formant frequency and sentiment score), fatigue (segmented TTR text complexity and sentiment score) and cognition (volume, harmonicity, spectral entropy, pause length standard deviation, and eyelid droop). The conversational, AI-enabled platform was feasible for extracting digital biomarkers in a sample of adults with depression. Taken together, our findings support the validity of previously discovered digital biomarkers and inform the development of personalized, remote monitoring models for mental health.

Keywords

biomarkers, depression, fatigue, cognition, multimodal sensors, large language models, computer vision, natural language processing

1. Introduction

Mental health, fatigue, and cognitive functioning are critical to overall well-being and operational performance, yet their assessment often relies on intermittent self-report or clinician-administered instruments. Such methods, while clinically validated, can be burdensome, subjective, and limited in their ability to capture dynamic changes over time [1, 2, 3, 4, 5, 6, 7, 8, 9]. Mobile health technology provides an alternative solution due to the ability of apps and sensors to collect high-fidelity and high-frequency data pertaining to activity, behavior, symptoms, and cognition [10, 11, 12]. With increasing demands for continuous and remote monitoring, especially in mental health, digital biomarkers have emerged as a promising avenue for objective and scalable assessment of psychological and cognitive states [13, 14, 15, 16].

Recent advances in computer vision, speech acoustics, and natural language processing have enabled the extraction of behavioral and physiological signals from everyday digital interactions [17, 18]. Multimodal sensing that integrates video, audio, and language data provides an opportunity to quantify indicators of affect, fatigue, and cognitive load to complement traditional instruments for assessment of mental health [19, 20]. Prior research has demonstrated associations between acoustic features (e.g., pitch variation [21, 22, 23, 24, 25, 26], jitter [21, 22, 27, 28, 29], and spectral harmonicity [28, 29, 30, 31]), facial expression dynamics (e.g., eye and pupil movements [32, 33, 34, 35, 36] and micro-expressions [37]), and linguistic markers (e.g., sentiment [31, 38, 39], complexity [31, 40], and emotional valence [31, 39, 41, 17, 42, 43, 44]) with symptoms of depression, stress, and cognitive decline. However, most

Joint Proceedings of the ACM Intelligent User Interfaces (IUI) Workshops 2026, March 23-26, 2026, Paphos, Cyprus
*Corresponding author.
matthew@okaya.me (M. So); michael.sobolev@cornell.edu (M. Sobolev); gregory@okaya.me (G. Menvielle)
0009-0005-1622-4249 (M. So); 0000-0002-8931-7682 (M. Sobolev)

existing work has been conducted in tightly controlled studies, and less is known about the feasibility of remote monitoring for collecting such data in naturalistic setting, such as mental health clinics [45].

Conversational AI agents, especially those powered by large language models (LLMs), introduce new opportunities for data collection in-situ [46]. By engaging users through dialogue, these systems can simultaneously elicit and record multimodal signals—speech, facial behavior, and language—while maintaining a familiar and low-burden interaction format. Such interfaces align with the increasing integration of conversational AI into mental health assessment and interventions, offering a pathway toward longitudinal, user-centered monitoring and treatment of mental health [47]. Nevertheless, empirical evaluation of these systems' feasibility and usability remains limited, particularly in diverse real-world populations.

The objective of this study was to evaluate the feasibility of a fully remote conversational AI agent, *Sanora*, for collecting analyzable multimodal data (video, audio, and text) in a population with mental health challenges, recruited from a women's health group in Australia. Specifically, we sought to determine whether participants would engage with the platform across multiple sessions and whether the system could reliably extract a comprehensive set of visual, acoustic, and linguistic features suitable for future digital biomarker modeling of depression, fatigue, and cognition. Finally, we present a case study for the design of risk score based on a model of extracted digital biomarkers to illustrate the future potential of the platform.

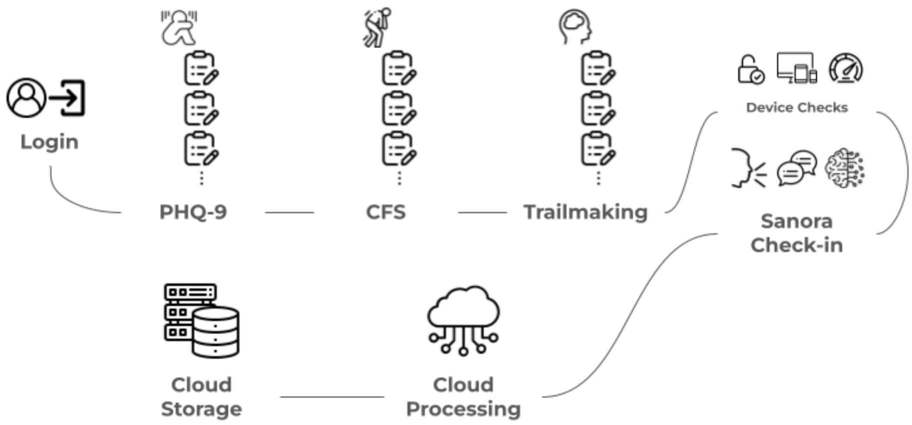

Figure 1: Illustration of the data collection workflow. The 5 steps for a successful sample collection include login, PHQ-9, CFS, Trailmaking, and Sanora Check-in.

2. Methods

2.1. Study Population

Participants were recruited from a women's well-being group in Australia (Pynk Health). Participation was entirely optional, and no data was collected except for email address which was required to access *Sanora* on the platform. The study was approved by Sterling Institutional Review Board (approval number 12300-GMenvielle), and all participants signed an Informed Consent Statement. A total of 10 unique users participated in the study, contributing 59 sessions (Median = 4.5; Min = 1; Max = 17).

Participants provided data during the period from February 19, 2025 to July 28, 2025, and received no compensation for their participation.

2.2. Procedure and Data Collection

Participants were sent weekly reminders to complete *Sanora* check-ins, but completion was entirely voluntary, and were given the option to view assessment results upon request. During the data collection period, a total of 24 weekly reminders were sent. Once users had successfully acknowledged the consent statement and registered for an Okaya account, users could complete a workflow as described in Fig. 1 [48]. During the check-in, users were first required to meet the device and browser requirements, grant permission to access the microphone and camera, and meet an upload speed requirement of 2mbps. Participants were then asked to complete 3 assessments - the Patient Health Questionnaire-9 (PHQ-9) [49], the Cancer Fatigue Scale (CFS) [50], and the Trail Making Test (TMT) [51]. It is important to note that users were allowed to use any browser or device that met the requirements, and took the assessment at any location, so it was difficult to control for environmental factors background noise and lighting.

Upon successful completion of all 3 assessments, users were presented with a text and voice prompt from an LLM (using OpenAI's *gpt-4o-mini* model with a custom system prompt, referred to as "*Sanora*"). Video was recorded continuously, and users could choose between webcam video, blurred video, a landmark representation of their face, an outline of their face, or a set of axes denoting the location of their face, designed to accommodate users with differing levels of comfort with being recorded. Though video was shown on the interface, raw video was not sent to our servers, instead, a set of facial landmarks was extracted via Google MediaPipe. These landmarks were automatically extracted with a proprietary machine learning model, and outputted 478 landmarks, each with an x, y, and z coordinate, as well as 52 "blendshapes" which each represented various facial expressions.

Sanora was designed to provide an engaging and comfortable environment by encouraging natural responses and supportive language. Table 1 summarizes small excerpts of the prompt we used and their associated purposes.

Upon clicking "Record" to record a response to *Sanora*, the user's microphone started recording and an automatic transcription service from AWS Transcribe started transcribing their speech, until the user stopped recording. Both the audio recording and transcribed text were sent back to our servers. *Sanora* then used the transcribed text to generate a response, which was then synthesized and played back to the user. This process continued for 4 turns (5 messages from *Sanora* and 4 responses from the participant).

2.3. Feature Extraction

From the landmarks, audio, and transcribed text, we extract 66 features motivated by the literature on digital biomarkers, mobile sensing, and mental health and cognitive assessment.

2.3.1. Visual Features

- blinks_per_s, blink_len - the number of detected blinks per second and average blink duration [52, 53, 54, 55, 56, 57]
- eyelid_droop, eyebrow_droop - how much the eyelids and eyebrows droop [43, 58]
- gaze_down_dist, gaze_x_dist, gaze_y_dist - how far the eyes gaze down or away from center [32, 33, 34, 35]
- eye_movements_per_s - the number of detected sharp eye movements per second [59]
- yawns_per_s, yawn_len - the number of detected yawns per second and average yawn duration [55, 60]
- mouth_curvature - average magnitude of smiling or frowning [34, 61, 62, 63, 58]
- movement_speed_measure - a subjective measure of average movement speed [61]
- affect_measure - overall amount of facial expression [37]

Purpose	Reasoning	Prompt Fragment
Persona & tone	Sanora should provide the user with a comfortable and safe environment to allow them to speak naturally.	"Be completely non-judgemental and come from a place of curiosity and interest."
Conversation boundaries	The user should not be allowed to speak about random topics Sanora is not designed to handle, and should be redirected in such situations. Additionally, Sanora shouldn't provide any diagnosis or treatment.	"Don't let the user change to a completely different topic."
		"Also, it is important you don't provide any recommendations or advice - your goal is purely just to assess and learn."
Engagement	The user should feel comfortable with speaking about their mental well-being and any possible stressors.	"Be completely non-judgemental and come from a place of curiosity and interest." "You should be an active listener and ask follow-up questions." "If the user doesn't provide engaging responses, use different questions or try to change the subject slightly if needed." "The user is very open to discussing emotions, feel free to ask deeper questions about their emotional state, feelings, and mental wellbeing." "The user prefers to avoid discussing emotions, focus on factual topics and behaviors rather than feelings. If emotions come up naturally, acknowledge them briefly but don't press for details."
Safety	For situations of immediate danger, the user should be pointed to a more suitable resource.	"In the case that they show any severe warning signs such as suicide, point them to professional help."
Continuity	The conversation should build upon previous conversations to encourage deeper and more personalized discussion.	"This is a transcript of your previous conversation: ... Quickly recap this previous conversation, then check up on their current wellbeing."

Table 1
Goals and design of the conversational agent Sanora

2.4. Audio Features

- pitch, pitch_std - mean and standard deviation of the frequency the user spoke at [21, 22, 23, 24, 25, 29, 30, 64, 65, 66]
- jitter, shimmer - variation in audio pitch and amplitude [21, 22, 28, 29, 64]
- vol, vol_std, vol_range - mean, standard deviation, and range of spoken volume detected from microphone [21, 25, 29, 67, 68, 69, 70, 64]
- crest_factor - ratio of peak volume to average volume
- volume_slope - trajectory of volume over time
- ealvi - Emotional Arousal Level Voice Index [71]

- timbre - timbral measurements of the audio, including the centroid, rolloff, flatness, affinity, sharpness, harmonicity, spectral tilt, spectral skewness, spectral kurtosis, and spectral entropy [22, 23, 27, 29, 30, 31, 64]
- formants - average F1, F2, F3, F4, and F5 formant frequencies [67, 72, 70]
- pauses - the number of pauses, as well as the mean and standard deviation of pauses [21, 24, 72, 73]

2.5. Text Features

- response_latency - average time it took to respond to Sanora [25, 74, 75]
- words_per_s - number of words spoken per second [25, 37, 72, 73, 74, 76, 77, 65]
- transcript_len - total number of words spoken [24, 73, 65]
- hesitations_per_s - number of hesitations (”uh,” ”um,” etc.) per second [77]
- complexity - various measurements for text complexity, including TTR, log TTR, segmented TTR, ARI, Coleman-Liau Index, Flesch-Kincaid Index, Gunning Index, and SMOG Index [31, 40]
- pauses - the number of pauses, as well as the mean and standard deviation of pauses [21, 24, 72, 73]
- emotion keyword proportions - proportion of words that are associated with a certain emotion, including anger, anticipation, disgust, fatigue, fear, joy, sadness, surprise, and trust [31, 39, 41, 17, 42, 43, 44]
- sentiment_score - positive or negative sentiment, scored by the VADER algorithm [31, 38, 39]

3. Results

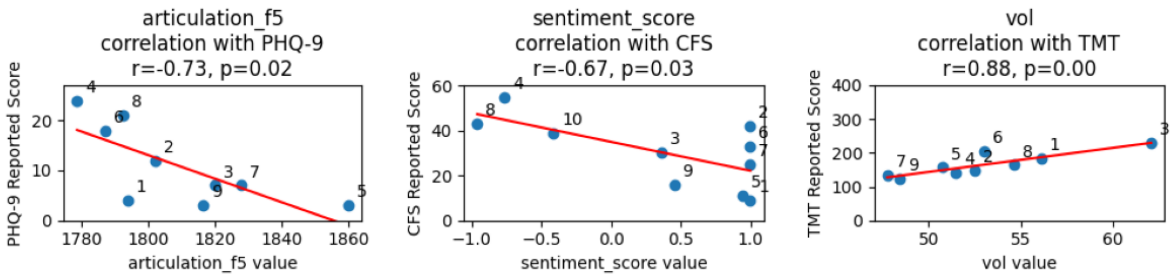

Figure 2: Most significant features for each measure and their associated correlation. For PHQ-9, F5 formant frequency is most significant, for CFS, it is text sentiment score, and for TMT, the most significant feature is volume.

3.1. Correlation Analysis

Pearson correlations and their associated significance are listed below. Though every participant was included in the dataset, we only analyzed the the first session for each participant. For each feature, outliers (defined as any measurement more than 2 standard deviations from the mean) were removed. The most significant features are plotted against the associated measures (Fig. 2).

3.1.1. Correlations for Depression

Exploratory correlation analyses revealed two significant associations between multimodal features and depressive symptom severity as measured by the PHQ-9. Acoustic and linguistic indicators showed the strongest relationships. Specifically, average F5 formant frequency (r = -0.73, p = 0.025) was negatively correlated with PHQ-9 scores, suggesting that a darker or more muffled vocal timbral quality may correspond to higher self-reported depression levels.

Textual sentiment, specifically the VADER sentiment score (r = -0.70, p = 0.025) also exhibited negative correlations with PHQ-9 scores. VADER is a sentiment algorithm that uses a social media-based lexicon to calculate a sentiment score either greater than 0.05 (positive sentiment), less than -0.05 (negative sentiment), or between 0.05 and -0.05 (neutral sentiment). These findings suggest that individuals reporting higher depressive symptoms might have speech content rated at a lower sentiment. Together, these relationships support the feasibility of deriving speech and language-based digital biomarkers of depression from naturalistic conversational data.

3.1.2. Correlations for Fatigue

Analysis of correlations between multimodal features and fatigue, as measured by the Cancer Fatigue Scale (CFS), identified two significant linguistic markers. The linguistic segmented TTR text complexity metric was inversely correlated with CFS scores (r = -0.67, p = 0.050), indicating that higher fatigue levels were associated with simpler or less complex speech production with less vocabulary diversity. Similarly to depression, the VADER sentiment score (r = -0.67, p = 0.034) also exhibited an inverse correlation with fatigue, suggesting that fatigued individuals might speak in a more negative manner. Together, these findings highlight the sensitivity of linguistic metrics to self-reported fatigue, supporting their potential use as digital biomarkers for real-time fatigue monitoring.

3.1.3. Correlations for Attention and Cognition

Exploratory analyses examining associations between multimodal features and cognitive performance, as measured by the Trail Making Test (TMT), identified the highest number of significant relationships. Among acoustic features, volume was positively correlated with TMT scores (r = 0.88, p = 0.0018), suggesting that users that speak louder might correspond to longer task completion times or lower cognitive efficiency. This is consistent with existing related cognition research such as ADHD, which found that higher speech volumes are tied to decreased inhibitory control [29]. Timbral harmonicity, which describes how periodic (leading to a stronger or clearer speech perception), was also positively correlated with TMT scores (r = 0.82, p = 0.0066). Lastly, spectral entropy was negatively correlated (r = -0.73, p = 0.025), suggesting that low spectral entropy that sounds clear, periodic, and tonal, might correspond to a deficiency in attention.

Among textual features, standard deviation of pause lengths exhibited an extremely high positive correlation with TMT scores (r = 0.97, p = 0.0064), which makes sense as individuals with high pause variability (exhibiting both short and long pauses), might correspond with difficulties with concentration.

Lastly, one visual feature, eyelid droop, was correlated with TMT scores (r = 0.78, p = 0.0076). In this case, a higher value means a more widely open eye, suggesting that widely open eyes correspond to higher TMT times. These results indicate that multimodal features from visual, audio, and textual sources derived from conversational speech may offer a feasible, low-burden proxy for cognitive performance monitoring in remote or operational settings.

3.1.4. The Effect of Outliers

To examine the effect of removing outliers when investigating correlations with such data, we calculated correlations with and without removing outliers. Though depression and fatigue measures were largely similar, TMT correlations were significantly different when outliers were not removed. Fig. 3 shows that 10 significant features are found as opposed to only 5 when outliers were removed. In these cases, outliers were so far from the rest of the dataset that a strong correlation was possible simply because the extreme values disproportionately influenced the regression line, artificially inflating the strength of the relationship between TMT performance and those features.

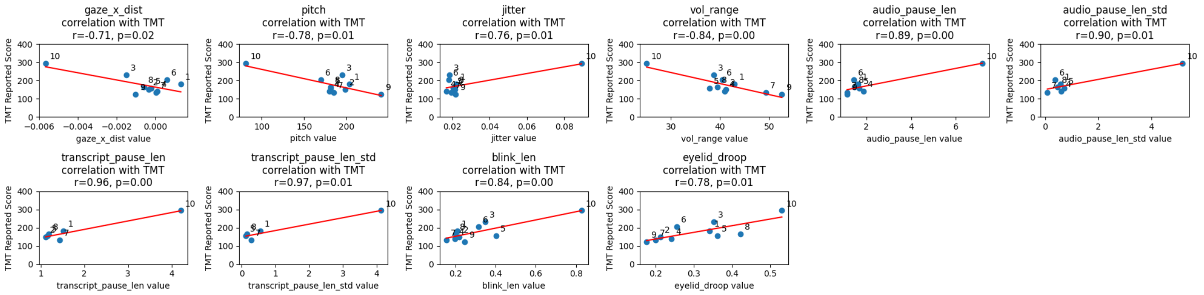

Figure 3: Significant features for the TMT and their associated correlations, with outliers not removed.

3.2. Qualitative Correlation Analysis

To complement the quantitative analyses, we conducted a qualitative review of the distribution of significant multimodal features across clinical domains. As shown in Table 2, depression-related features were primarily linguistic and acoustic in nature, with significant visual markers identified only for questions 4, 5, and 7. Interestingly, question 5 (poor appetite) exhibited the most significant features, with 3 visual, 2 audio, and 2 text features significantly correlated. In contrast, question 9 (suicide and self-harm) exhibited no significant features.

Fatigue, as measured by the CFS, was primarily associated with linguistic features, with subscales including the physical and affective subscales associating with the audio modality as well. No visual features were found to have significant correlations, including individual subscales.

Cognitive performance exhibited a broad feature coverage, encompassing one visual, three acoustic, and one linguistic variable.

	Feature Categories		
Measure	Visual (# features)	Audio (# features)	Text (# features)
PHQ-9	0	1	1
Q1	0	1	0
Q2	0	1	1
Q3	0	1	1
Q4	1	2	0
Q5	3	2	2
Q6	0	1	1
Q7	2	0	1
Q8	0	2	0
Q9	0	0	0
CFS	0	0	2
Physical	0	1	2
Affective	0	1	3
Cognitive	0	0	0
TMT	1	3	1
Part A	0	1	1
Part B	1	1	1

Table 2
Count of number of significant features for each clinical measure, individual questions, and subscales per feature type

3.3. Risk Score Case Study

To explore the feasibility of individualized monitoring, we demonstrate case studies on participants who completed multiple sessions with *Sanora*. For each individual, a composite "risk score" was calculated by standardizing the significant multimodal features (z-scoring within participant) and summing up the deviations from the mean, with directionality determined by the feature's correlation with the clinical measure. This approach provided an interpretable, participant-specific index of deviation from baseline functioning across time.

Longitudinal visualization of these risk scores revealed dynamic fluctuations in depression, fatigue, and cognitive indices across sessions. Although exploratory, these case studies illustrate how multimodal conversational data can support within-person anomaly detection and temporal tracking of psychological and physiological states. The results suggest that individualized, feature-based composite scores may offer a viable framework for early identification of changes in mood, fatigue, or cognitive function in remote monitoring contexts.

Figures for the 2 participants with the most sessions are displayed for depression, fatigue, and cognition in Fig. 4.

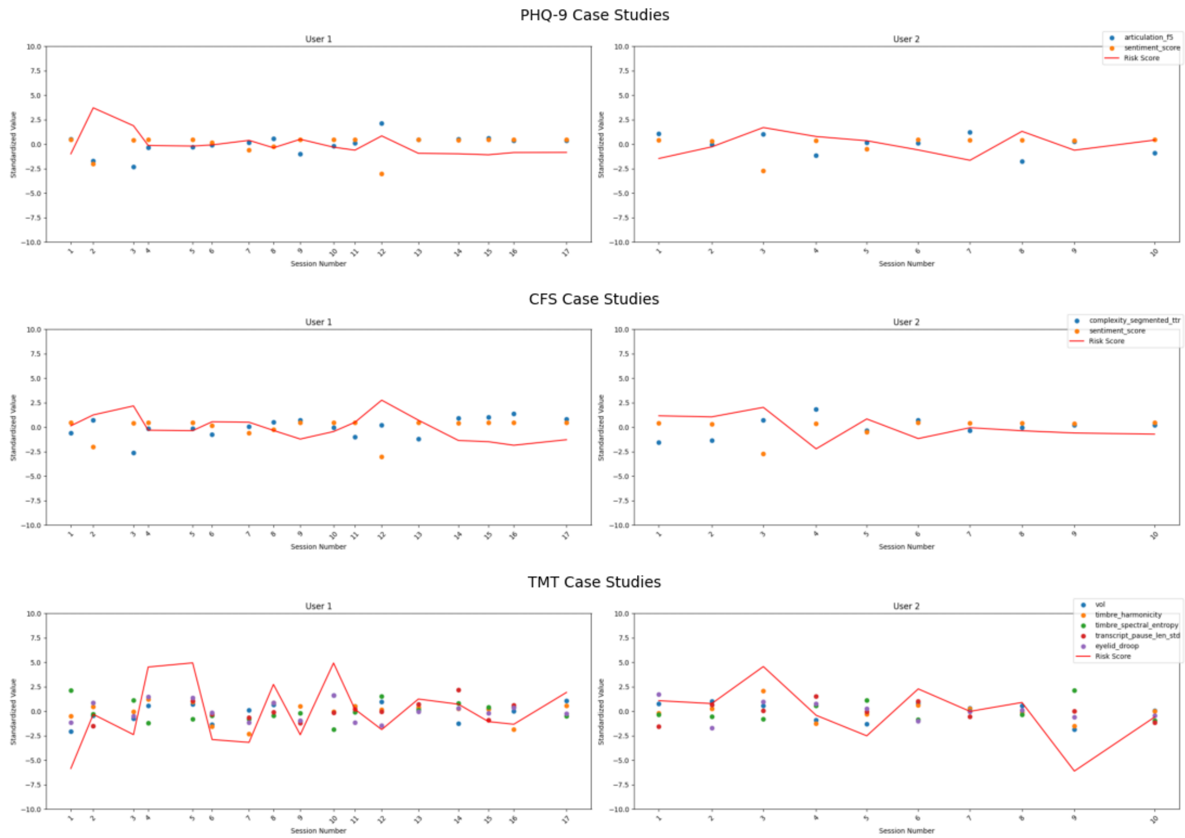

Figure 4: Risk score case study for depression, fatigue, and cognition.

4. Discussion

We found several multimodal features that correlated with depression, fatigue, and cognition, some of which extended beyond patterns reported in prior literature. As expected, depression was characterized by a more muffled timbre and more negative sentiment, consistent with psychomotor slowing and affective flattening. Interestingly, question 5 (poor appetite or overeating) exhibited the most significant features, while question 9 (thoughts that you would be better off dead or of hurting yourself in some way) exhibited no significant features. This suggests a possible difficulty in using the biomarkers Sanora collects as an indicator of suicidal ideation or risk, in addition to the fact that depression is a multi-faceted construct that should not be treated as a single monolith. Fatigue demonstrated a similar pattern, including reduced complexity and increased negative sentiment. Notably, complexity and sentiment, originally hypothesized as depression-related markers, were also significantly associated with fatigue, suggesting that they may reflect shared affective mechanisms such as diminished emotional expressivity and lexical complexity. It is notable that fatigue, including the individual subscales, had no significant correlations with visual features, suggesting a possibility of fatigue not manifesting itself in visual signs. Attention and cognition, as measured by the TMT, demonstrated the broadest multimodal signature across all three domains, including increased volume and harmonicity, decreased spectral entropy, high pause variability, and low eyelid droop. Similarly, audio measurements including pause variability, a feature previously linked to depressive and fatigued speech, was found to correlate with cognitive performance, indicating that subtle fluctuations in various vocal quantities may also capture attentional or executive variability. Additionally, eyelid droop was an unexpected visual indicator for attention, when it was meant as an indicator for fatigue, though in the opposite correlation direction. Lastly, our outlier analysis provided a more realistic interpretation of our results, especially with the TMT. It is worth noting that, since TMT is much more likely to be affected by concentration compared to assessments such as the PHQ-9 and CFS, this may introduce some noise to the TMT (but could also partly act as an indicator of cognition). Together, these findings highlight both the replicability of established digital biomarkers and the emergence of cross-domain indicators that may bridge affective, cognitive, and fatigue-related processes.

5. Conclusion

The findings from the case study align closely with the long-term vision of the conversational AI agent we developed, *Sanora*. The observed fluctuations in individualized risk scores and their correspondence with self-reported symptom changes and objective cognitive assessment demonstrate the agent's potential for longitudinal, within-person monitoring of mental health. Going forward, *Sanora* design will emphasizes personalized baselines and interprets deviations relative to each individual's unique behavioral and physiological profile. This will allow early identification of shifts in emotional, cognitive, or fatigue states without relying on additional assessments. This trajectory aligns with the goal of advancing ethically grounded, data-secure, and personalized digital biomarkers for proactive mental health assessment and treatment. It is important to note that, due to our low sample size, we are not yet able to make claims around validation and interpretability of our results. However, as digital biomarkers move closer to clinical integration, continued validation across diverse populations and conditions will be essential to ensure generalizability, interpretability, and equity [78, 79].

This study has several limitations that should be considered when interpreting the findings. First, the sample size was small and limited due to our focus on a single, small clinic, as well as the optionality of data collection, constraining generalizability and the ability to model inter-individual variability. Second, although validated instruments such as the PHQ-9, CFS, and TMT were used for comparison, these measures may not fully capture the temporal dynamics or subtle within-person fluctuations that multimodal digital biomarkers are designed to detect. Traditional self-report and cognitive assessments tend to be relatively stable, which may underestimate moment-to-moment variability in affect, fatigue, and cognition. Future studies should incorporate higher-resolution clinical measures—such as ecological momentary assessment (EMA) or brief daily self-reports—that provide richer temporal data and enable stronger coupling between behavioral features and self-reported states [1, 2, 3]. Future studies might also complement this by incorporating other more continuous data sources such as wearables [80]. Integrating these temporally aligned measures would allow for more precise modeling of intra-individual change and enhance the ecological validity of the derived digital biomarkers. Finally, while the AI agent *Sanora* was deemed feasible in the correct context, further validation with larger, more diverse samples and extended longitudinal follow-up is necessary for predictive modeling and design of personalized remote monitoring systems for mental health.

Acknowledgments

We would like to thank Sara Ghandehari, MD, MCSO, FACHE, as well as Nick Uram, Psy.D for their invaluable contributions towards the development of the Okaya platform. We are also grateful to Will Jackson for his assistance in literature review.

Declaration on Generative AI

The authors have not employed any Generative AI tools in the writing of this paper.

References

[1] M. J. Sliwinski, J. A. Mogle, J. Hyun, E. Munoz, J. M. Smyth, R. B. Lipton, Reliability and Validity of Ambulatory Cognitive Assessments, Assessment 25 (2018) 14–30. URL: http://journals.sagepub.com/doi/10.1177/1073191116643164. doi:10.1177/1073191116643164.

[2] M. J. Katz, C. Wang, C. O. Nester, C. A. Derby, M. E. Zimmerman, R. B. Lipton, M. J. Sliwinski, L. A. Rabin, T-MoCA: A valid phone screen for cognitive impairment in diverse community samples, Alzheimer's & Dementia: Diagnosis, Assessment & Disease Monitoring 13 (2021) e12144. URL: https://alz-journals.onlinelibrary.wiley.com/doi/10.1002/dad2.12144. doi:10.1002/dad2.12144.

[3] S. Singh, R. Strong, I. Xu, L. M. Fonseca, Z. Hawks, E. Grinspoon, L. Jung, F. Li, R. S. Weinstock, M. J. Sliwinski, N. S. Chaytor, L. T. Germine, Ecological Momentary Assessment of Cognition in Clinical and Community Samples: Reliability and Validity Study, Journal of Medical Internet Research 25 (2023) e45028. URL: https://www.jmir.org/2023/1/e45028. doi:10.2196/45028.

[4] R. K. McHugh, E. Behar, Readability of self-report measures of depression and anxiety., Journal of Consulting and Clinical Psychology 77 (2009) 1100–1112. URL: https://doi.apa.org/doi/10.1037/a0017124. doi:10.1037/a0017124.

[5] R. K. McHugh, E. Behar, Reply to "Further issues in determining the readability of self-report items: Comment on McHugh and Behar (2009)"., Journal of Consulting and Clinical Psychology 80 (2012) 1121–1122. URL: https://doi.apa.org/doi/10.1037/a0030455. doi:10.1037/a0030455.

[6] J. A. Schinka, Further issues in determining the readability of self-report items: Comment on McHugh and Behar (2009)., Journal of Consulting and Clinical Psychology 80 (2012) 952–955. URL: https://doi.apa.org/doi/10.1037/a0029928. doi:10.1037/a0029928.

[7] B. Levis, A. Benedetti, J. P. Ioannidis, Y. Sun, Z. Negeri, C. He, Y. Wu, A. Krishnan, P. M. Bhandari, D. Neupane, M. Imran, D. B. Rice, K. E. Riehm, N. Saadat, M. Azar, J. Boruff, P. Cuijpers, S. Gilbody, L. A. Kloda, D. McMillan, S. B. Patten, I. Shrier, R. C. Ziegelstein, S. H. Alamri, D. Amtmann, L. Ayalon, H. R. Baradaran, A. Beraldi, C. N. Bernstein, A. Bhana, C. H. Bombardier, G. Carter, M. H. Chagas, D. Chibanda, K. Clover, Y. Conwell, C. Diez-Quevedo, J. R. Fann, F. H. Fischer, L. Gholizadeh, L. J. Gibson, E. P. Green, C. G. Greeno, B. J. Hall, E. E. Haroz, K. Ismail, N. Jetté, M. E. Khamseh, Y. Kwan, M. A. Lara, S.-I. Liu, S. R. Loureiro, B. Löwe, R. A. Marrie, L. Marsh, A. McGuire, K. Muramatsu, L. Navarrete, F. L. Osório, I. Petersen, A. Picardi, S. L. Pugh, T. J. Quinn, A. G. Rooney, E. H. Shinn, A. Sidebottom, L. Spangenberg, P. L. L. Tan, M. Taylor-Rowan, A. Turner, H. C. Van Weert, P. A. Vöhringer, L. I. Wagner, J. White, K. Winkley, B. D. Thombs, Patient Health Questionnaire-9 scores do not accurately estimate depression prevalence: individual participant data meta-analysis, Journal of Clinical Epidemiology 122 (2020) 115–128.e1. URL: https://linkinghub.elsevier.com/retrieve/pii/S0895435619307358. doi:10.1016/j.jclinepi.2020.02.002.

[8] S. Ma, L. Kang, X. Guo, H. Liu, L. Yao, H. Bai, C. Chen, M. Hu, L. Du, H. Du, C. Ai, F. Wang, G. Wang, R. Li, Z. Liu, Discrepancies between self-rated depression and observed depression severity: The effects of personality and dysfunctional attitudes, General Hospital Psychiatry 70

(2021) 25–30. URL: https://linkinghub.elsevier.com/retrieve/pii/S0163834320301687. doi:10.1016/j.genhosppsych.2020.11.016.

[9] T. Inoue, T. Tanaka, S. Nakagawa, Y. Nakato, R. Kameyama, S. Boku, H. Toda, T. Kurita, T. Koyama, Utility and limitations of PHQ-9 in a clinic specializing in psychiatric care, BMC Psychiatry 12 (2012) 73. URL: http://bmcpsychiatry.biomedcentral.com/articles/10.1186/1471-244X-12-73. doi:10.1186/1471-244X-12-73.

[10] I. Sim, Mobile Devices and Health, The New England Journal of Medicine 381 (2019) 956–968. doi:10.1056/NEJMra1806949.

[11] T. R. Insel, Digital Phenotyping: Technology for a New Science of Behavior, JAMA 318 (2017) 1215–1216. doi:10.1001/jama.2017.11295.

[12] D. C. Mohr, M. Zhang, S. M. Schueller, Personal Sensing: Understanding Mental Health Using Ubiquitous Sensors and Machine Learning, Annual Review of Clinical Psychology 13 (2017) 23–47. doi:10.1146/annurev-clinpsy-032816-044949.

[13] A. Coravos, S. Khozin, K. D. Mandl, Developing and adopting safe and effective digital biomarkers to improve patient outcomes, npj Digital Medicine 2 (2019) 14. URL: https://www.nature.com/articles/s41746-019-0090-4. doi:10.1038/s41746-019-0090-4.

[14] P. Dagum, Digital biomarkers of cognitive function, npj Digital Medicine 1 (2018) 10. URL: https://www.nature.com/articles/s41746-018-0018-4. doi:10.1038/s41746-018-0018-4.

[15] N. C. Jacobson, H. Weingarden, S. Wilhelm, Digital biomarkers of mood disorders and symptom change, npj Digital Medicine 2 (2019) 3. URL: https://www.nature.com/articles/s41746-019-0078-0. doi:10.1038/s41746-019-0078-0.

[16] M. Sobolev, B. T. Gullapalli, T. Rahman, Advancing the science of digital biomarkers, in: Proceedings of the 2021 Workshop on Future of Digital Biomarkers, ACM, Virtual Event Wisconsin, 2021, pp. 1–2. URL: https://dl.acm.org/doi/10.1145/3469266.3473711. doi:10.1145/3469266.3473711.

[17] J. Han, Y. Feng, N. Li, L. Feng, L. Xiao, X. Zhu, G. Wang, Correlation Between Word Frequency and 17 Items of Hamilton Scale in Major Depressive Disorder, Frontiers in Psychiatry 13 (2022) 902873. URL: https://www.frontiersin.org/articles/10.3389/fpsyt.2022.902873/full. doi:10.3389/fpsyt.2022.902873.

[18] M. R. Islam, D. Sarker, M. T. H. Sadi, L. H. Bayzid, M. Kabiruzzaman, S. Hossain, Facial Behavior-Based Depression Detection with Bi-LSTM and Evaluation via Universal and Hybrid Methods, in: 2025 International Conference on Activity and Behavior Computing (ABC), IEEE, Al Ain, United Arab Emirates, 2025, pp. 1–8. URL: https://ieeexplore.ieee.org/document/11118568/. doi:10.1109/ABC64332.2025.11118568.

[19] E. Garcia-Ceja, M. Riegler, T. Nordgreen, P. Jakobsen, K. J. Oedegaard, J. Tørresen, Mental health monitoring with multimodal sensing and machine learning: A survey, Pervasive and Mobile Computing 51 (2018) 1–26. URL: https://linkinghub.elsevier.com/retrieve/pii/S1574119217305692. doi:10.1016/j.pmcj.2018.09.003.

[20] T. R. Insel, Digital Phenotyping: Technology for a New Science of Behavior, JAMA 318 (2017) 1215–1216. doi:10.1001/jama.2017.11295.

[21] A. König, J. Tröger, E. Mallick, M. Mina, N. Linz, C. Wagnon, J. Karbach, C. Kuhn, J. Peter, Detecting subtle signs of depression with automated speech analysis in a non-clinical sample, BMC Psychiatry 22 (2022) 830. URL: https://bmcpsychiatry.biomedcentral.com/articles/10.1186/s12888-022-04475-0. doi:10.1186/s12888-022-04475-0.

[22] W. J. Silva, L. Lopes, M. K. C. Galdino, A. A. Almeida, Voice Acoustic Parameters as Predictors of Depression, Journal of Voice 38 (2024) 77–85. URL: https://linkinghub.elsevier.com/retrieve/pii/S0892199721002058. doi:10.1016/j.jvoice.2021.06.018.

[23] D. Shin, W. Cho, C. Park, S. Rhee, M. Kim, H. Lee, N. Kim, Y. Ahn, Detection of Minor and Major Depression through Voice as a Biomarker Using Machine Learning, Journal of Clinical Medicine 10 (2021) 3046. URL: https://www.mdpi.com/2077-0383/10/14/3046. doi:10.3390/jcm10143046.

[24] M. Alpert, E. R. Pouget, R. R. Silva, Reflections of depression in acoustic measures of the patient's speech, Journal of Affective Disorders 66 (2001) 59–69. URL: https://linkinghub.elsevier.com/retrieve/pii/S0165032700003359. doi:10.1016/S0165-0327(00)00335-9.

[25] A. S. Cohen, Z. Rodriguez, M. Opler, B. Kirkpatrick, S. Milanovic, D. Piacentino, S. T. Szabo, S. Tomioka, A. Ogirala, K. S. Koblan, J. S. Siegel, S. Hopkins, Evaluating speech latencies during structured psychiatric interviews as an automated objective measure of psychomotor slowing, Psychiatry Research 340 (2024) 116104. URL: https://linkinghub.elsevier.com/retrieve/pii/S0165178124003895. doi:10.1016/j.psychres.2024.116104.

[26] A. P. Vogel, J. Fletcher, P. Maruff, Acoustic analysis of the effects of sustained wakefulness on speech, The Journal of the Acoustical Society of America 128 (2010) 3747–3756. URL: https://pubs.aip.org/jasa/article/128/6/3747/904268/Acoustic-analysis-of-the-effects-of-sustained. doi:10.1121/1.3506349.

[27] A. Ozdas, R. Shiavi, S. Silverman, M. Silverman, D. Wilkes, Investigation of Vocal Jitter and Glottal Flow Spectrum as Possible Cues for Depression and Near-Term Suicidal Risk, IEEE Transactions on Biomedical Engineering 51 (2004) 1530–1540. URL: http://ieeexplore.ieee.org/document/1325813/. doi:10.1109/TBME.2004.827544.

[28] X. Gao, K. Ma, H. Yang, K. Wang, B. Fu, Y. Zhu, X. She, B. Cui, A rapid, non-invasive method for fatigue detection based on voice information, Frontiers in Cell and Developmental Biology 10 (2022) 994001. URL: https://www.frontiersin.org/articles/10.3389/fcell.2022.994001/full. doi:10.3389/fcell.2022.994001.

[29] A.-L. Hamdan, R. Deeb, A. Sibai, C. Rameh, H. Rifai, J. Fayyad, Vocal Characteristics in Children With Attention Deficit Hyperactivity Disorder, Journal of Voice 23 (2009) 190–194. URL: https://linkinghub.elsevier.com/retrieve/pii/S0892199707001233. doi:10.1016/j.jvoice.2007.09.004.

[30] Q. Zhao, H.-Z. Fan, Y.-L. Li, L. Liu, Y.-X. Wu, Y.-L. Zhao, Z.-X. Tian, Z.-R. Wang, Y.-L. Tan, S.-P. Tan, Vocal Acoustic Features as Potential Biomarkers for Identifying/Diagnosing Depression: A Cross-Sectional Study, Frontiers in Psychiatry 13 (2022) 815678. URL: https://www.frontiersin.org/articles/10.3389/fpsyt.2022.815678/full. doi:10.3389/fpsyt.2022.815678.

[31] 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.

[32] M. K. Hinchliffe, M. H. Lancashire, F. J. Roberts, Eye-Contact and Depression: A Preliminary Report, British Journal of Psychiatry 117 (1970) 571–572. URL: https://www.cambridge.org/core/product/identifier/S0007125000194310/type/journal_article. doi:10.1192/bjp.117.540.571.

[33] T. Suslow, D. Hoepfel, A. Kersting, C. M. Bodenschatz, Depressive symptoms and visual attention to others' eyes in healthy individuals, BMC Psychiatry 24 (2024) 184. URL: https://bmcpsychiatry.biomedcentral.com/articles/10.1186/s12888-024-05633-2. doi:10.1186/s12888-024-05633-2.

[34] J. T. Fiquer, R. A. Moreno, A. R. Brunoni, V. B. Barros, F. Fernandes, C. Gorenstein, What is the nonverbal communication of depression? Assessing expressive differences between depressive patients and healthy volunteers during clinical interviews, Journal of Affective Disorders 238 (2018) 636–644. URL: https://linkinghub.elsevier.com/retrieve/pii/S0165032717321456. doi:10.1016/j.jad.2018.05.071.

[35] P. Waxer, Nonverbal cues for depression., Journal of Abnormal Psychology 83 (1974) 319–322. URL: https://doi.apa.org/doi/10.1037/h0036706. doi:10.1037/h0036706.

[36] R. Islam, S. W. Bae, PupilSense: Detection of Depressive Episodes through Pupillary Response in the Wild, in: 2024 International Conference on Activity and Behavior Computing (ABC), IEEE, Oita / Kitakyushu, Japan, 2024, pp. 01–13. URL: https://ieeexplore.ieee.org/document/10652166/. doi:10.1109/ABC61795.2024.10652166.

[37] J. S. Buyukdura, S. M. McClintock, P. E. Croarkin, Psychomotor retardation in depression: Biological underpinnings, measurement, and treatment, Progress in Neuro-Psychopharmacology and Biological Psychiatry 35 (2011) 395–409. URL: https://linkinghub.elsevier.com/retrieve/pii/S0278584610004057. doi:10.1016/j.pnpbp.2010.10.019.

[38] B. Hadžić, J. Ohse, M. Alkostantini, N. Peperkorn, A. Yorita, T. Weber, N. Kubota, Y. Shiban, M. Rätsch, Emotional Dynamics in Semi-Clinical Settings: Speech Emotion Recognition in Depression-Related Interviews:, in: Proceedings of the 11th International Conference on Infor-

mation and Communication Technologies for Ageing Well and e-Health, SCITEPRESS - Science and Technology Publications, Porto, Portugal, 2025, pp. 104–113. URL: https://www.scitepress.org/DigitalLibrary/Link.aspx?doi=10.5220/0013415700003938. doi:10.5220/0013415700003938.

[39] J. Rottenberg, J. J. Gross, I. H. Gotlib, Emotion Context Insensitivity in Major Depressive Disorder., Journal of Abnormal Psychology 114 (2005) 627–639. URL: https://doi.apa.org/doi/10.1037/0021-843X.114.4.627. doi:10.1037/0021-843X.114.4.627.

[40] R. N. Trifu, B. Nemeș, D. C. Herta, C. Bodea-Hategan, D. A. Talaș, H. Coman, Linguistic markers for major depressive disorder: a cross-sectional study using an automated procedure, Frontiers in Psychology 15 (2024) 1355734. URL: https://www.frontiersin.org/articles/10.3389/fpsyg.2024.1355734/full. doi:10.3389/fpsyg.2024.1355734.

[41] L. Hartnagel, U. W. Ebner-Priemer, J. C. Foo, F. Streit, S. H. Witt, J. Frank, M. F. Limberger, A. B. Horn, M. Gilles, M. Rietschel, L. Sirignano, Linguistic style as a digital marker for depression severity: An ambulatory assessment pilot study in patients with depressive disorder undergoing sleep deprivation therapy, Acta Psychiatrica Scandinavica 151 (2025) 348–357. URL: https://onlinelibrary.wiley.com/doi/10.1111/acps.13726. doi:10.1111/acps.13726.

[42] S. Shinohara, M. Nakamura, Y. Omiya, M. Higuchi, N. Hagiwara, S. Mitsuyoshi, H. Toda, T. Saito, M. Tanichi, A. Yoshino, S. Tokuno, Depressive Mood Assessment Method Based on Emotion Level Derived from Voice: Comparison of Voice Features of Individuals with Major Depressive Disorders and Healthy Controls, International Journal of Environmental Research and Public Health 18 (2021) 5435. URL: https://www.mdpi.com/1660-4601/18/10/5435. doi:10.3390/ijerph18105435.

[43] J.-G. Gehricke, D. Shapiro, Reduced facial expression and social context in major depression: discrepancies between facial muscle activity and self-reported emotion, Psychiatry Research 95 (2000) 157–167. URL: https://linkinghub.elsevier.com/retrieve/pii/S0165178100001682. doi:10.1016/S0165-1781(00)00168-2.

[44] L. A. Cariola, L. B. Sheeber, N. Allen, M. Bilalpur, T. Bird, S. Hinduja, L.-P. Morency, J. F. Cohn, Language use in depressed and non-depressed mothers and their adolescent offspring, Journal of Affective Disorders 366 (2024) 290–299. URL: https://linkinghub.elsevier.com/retrieve/pii/S0165032724013570. doi:10.1016/j.jad.2024.08.131.

[45] S. L. Connolly, E. Kuhn, K. Possemato, J. Torous, Digital Clinics and Mobile Technology Implementation for Mental Health Care, Current Psychiatry Reports 23 (2021) 38. URL: https://link.springer.com/10.1007/s11920-021-01254-8. doi:10.1007/s11920-021-01254-8.

[46] J. Torous, E. J. Topol, Assessing generative artificial intelligence for mental health, The Lancet 406 (2025) 683. URL: https://linkinghub.elsevier.com/retrieve/pii/S0140673625012371. doi:10.1016/S0140-6736(25)01237-1.

[47] M. V. Heinz, D. M. Mackin, B. M. Trudeau, S. Bhattacharya, Y. Wang, H. A. Banta, A. D. Jewett, A. J. Salzhauer, T. Z. Griffin, N. C. Jacobson, Randomized Trial of a Generative AI Chatbot for Mental Health Treatment, NEJM AI 2 (2025). URL: https://ai.nejm.org/doi/10.1056/AIoa2400802. doi:10.1056/AIoa2400802.

[48] M. So, M. Sobolev, G. Menvielle, Sanora: A Conversational AI Agent for Multimodal Digital Biomarkers of Mental Health, ???? URL: https://www.techrxiv.org/users/1027431/articles/1387167-sanora-a-conversational-ai-agent-for-multimodal-digital-biomarkers-of-mental-health.

[49] K. Kroenke, R. L. Spitzer, J. B. W. Williams, The PHQ-9: Validity of a brief depression severity measure, Journal of General Internal Medicine 16 (2001) 606–613. URL: http://link.springer.com/10.1046/j.1525-1497.2001.016009606.x. doi:10.1046/j.1525-1497.2001.016009606.x.

[50] T. Okuyama, T. Akechi, A. Kugaya, H. Okamura, Y. Shima, M. Maruguchi, T. Hosaka, Y. Uchitomi, Development and Validation of the Cancer Fatigue Scale, Journal of Pain and Symptom Management 19 (2000) 5–14. URL: https://linkinghub.elsevier.com/retrieve/pii/S0885392499001384. doi:10.1016/S0885-3924(99)00138-4.

[51] J. Dahmen, D. Cook, R. Fellows, M. Schmitter-Edgecombe, An analysis of a digital variant of the Trail Making Test using machine learning techniques, Technology and Health Care 25 (2017) 251–264. URL: https://journals.sagepub.com/doi/full/10.3233/THC-161274. doi:10.3233/THC-161274.

[52] R. Zargari Marandi, P. Madeleine, Omland, N. Vuillerme, A. Samani, Eye movement char-

acteristics reflected fatigue development in both young and elderly individuals, Scientific Reports 8 (2018) 13148. URL: https://www.nature.com/articles/s41598-018-31577-1. doi:10.1038/s41598-018-31577-1.

[53] J. A. Stern, D. Boyer, D. Schroeder, Blink Rate: A Possible Measure of Fatigue, Human Factors: The Journal of the Human Factors and Ergonomics Society 36 (1994) 285–297. URL: https://journals.sagepub.com/doi/10.1177/001872089403600209. doi:10.1177/001872089403600209.

[54] Z. Yin, B. Liu, D. Hao, L. Yang, Y. Feng, Evaluation of VDT-Induced Visual Fatigue by Automatic Detection of Blink Features, Sensors 22 (2022) 916. URL: https://www.mdpi.com/1424-8220/22/3/916. doi:10.3390/s22030916.

[55] F. Hasan, A. Kashevnik, State-of-the-Art Analysis of Modern Drowsiness Detection Algorithms Based on Computer Vision, in: 2021 29th Conference of Open Innovations Association (FRUCT), IEEE, Tampere, Finland, 2021, pp. 141–149. URL: https://ieeexplore.ieee.org/document/9435480/. doi:10.23919/FRUCT52173.2021.9435480.

[56] Y. Yamada, M. Kobayashi, Detecting mental fatigue from eye-tracking data gathered while watching video: Evaluation in younger and older adults, Artificial Intelligence in Medicine 91 (2018) 39–48. URL: https://linkinghub.elsevier.com/retrieve/pii/S0933365717306140. doi:10.1016/j.artmed.2018.06.005.

[57] A. Magliacano, S. Fiorenza, A. Estraneo, L. Trojano, Eye blink rate increases as a function of cognitive load during an auditory oddball paradigm, Neuroscience Letters 736 (2020) 135293. URL: https://linkinghub.elsevier.com/retrieve/pii/S0304394020305632. doi:10.1016/j.neulet.2020.135293.

[58] T. Sundelin, M. Lekander, G. Kecklund, E. J. W. Van Someren, A. Olsson, J. Axelsson, Cues of Fatigue: Effects of Sleep Deprivation on Facial Appearance, Sleep 36 (2013) 1355–1360. URL: https://academic.oup.com/sleep/article/36/9/1355/2453883. doi:10.5665/sleep.2964.

[59] B. Fischer, B. Breitmeyer, Mechanisms of visual attention revealed by saccadic eye movements, Neuropsychologia 25 (1987) 73–83. URL: https://linkinghub.elsevier.com/retrieve/pii/0028393287900443. doi:10.1016/0028-3932(87)90044-3.

[60] A. G. Guggisberg, J. Mathis, A. Schnider, C. W. Hess, Why do we yawn?, Neuroscience & Biobehavioral Reviews 34 (2010) 1267–1276. URL: https://linkinghub.elsevier.com/retrieve/pii/S0149763410000825. doi:10.1016/j.neubiorev.2010.03.008.

[61] J. M. Girard, J. F. Cohn, M. H. Mahoor, S. M. Mavadati, Z. Hammal, D. P. Rosenwald, Nonverbal social withdrawal in depression: Evidence from manual and automatic analyses, Image and Vision Computing 32 (2014) 641–647. URL: https://linkinghub.elsevier.com/retrieve/pii/S0262885613001790. doi:10.1016/j.imavis.2013.12.007.

[62] S. Scherer, G. Stratou, M. Mahmoud, J. Boberg, J. Gratch, A. Rizzo, L.-P. Morency, Automatic behavior descriptors for psychological disorder analysis, in: 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), IEEE, Shanghai, China, 2013, pp. 1–8. URL: http://ieeexplore.ieee.org/document/6553789/. doi:10.1109/FG.2013.6553789.

[63] J. M. Girard, J. F. Cohn, M. H. Mahoor, S. Mavadati, D. P. Rosenwald, Social risk and depression: Evidence from manual and automatic facial expression analysis, in: 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), IEEE, Shanghai, China, 2013, pp. 1–8. URL: http://ieeexplore.ieee.org/document/6553748/. doi:10.1109/FG.2013.6553748.

[64] X. Gao, K. Ma, H. Yang, K. Wang, B. Fu, Y. Zhu, X. She, B. Cui, A rapid, non-invasive method for fatigue detection based on voice information, Frontiers in Cell and Developmental Biology 10 (2022) 994001. URL: https://www.frontiersin.org/articles/10.3389/fcell.2022.994001/full. doi:10.3389/fcell.2022.994001.

[65] A. P. Vogel, J. Fletcher, P. Maruff, Acoustic analysis of the effects of sustained wakefulness on speech, The Journal of the Acoustical Society of America 128 (2010) 3747–3756. URL: https://pubs.aip.org/jasa/article/128/6/3747/904268/Acoustic-analysis-of-the-effects-of-sustained. doi:10.1121/1.3506349.

[66] Y. Bloch, S. Aviram, R. Neeman, Y. Braw, U. Nitzan, H. Maoz, A. Mimouni-Bloch, Methylphenidate

mediated change in prosody is specific to the performance of a cognitive task in female adult ADHD patients, The World Journal of Biological Psychiatry 16 (2015) 635–639. URL: https://www.tandfonline.com/doi/full/10.3109/15622975.2015.1036115. doi:10.3109/15622975.2015.1036115.

[67] F. Menne, F. Dörr, J. Schräder, J. Tröger, U. Habel, A. König, L. Wagels, The voice of depression: speech features as biomarkers for major depressive disorder, BMC Psychiatry 24 (2024) 794. URL: https://bmcpsychiatry.biomedcentral.com/articles/10.1186/s12888-024-06253-6. doi:10.1186/s12888-024-06253-6.

[68] Y. Wang, L. Liang, Z. Zhang, X. Xu, R. Liu, H. Fang, R. Zhang, Y. Wei, Z. Liu, R. Zhu, X. Zhang, F. Wang, Fast and accurate assessment of depression based on voice acoustic features: a cross-sectional and longitudinal study, Frontiers in Psychiatry 14 (2023) 1195276. URL: https://www.frontiersin.org/articles/10.3389/fpsyt.2023.1195276/full. doi:10.3389/fpsyt.2023.1195276.

[69] N. Cummins, V. Sethu, J. Epps, J. Krajewski, Probabilistic acoustic volume analysis for speech affected by depression, in: Interspeech 2014, ISCA, 2014, pp. 1238–1242. URL: https://www.isca-archive.org/interspeech_2014/cummins14_interspeech.html. doi:10.21437/Interspeech.2014-311.

[70] J. Krajewski, U. Trutschel, M. Golz, D. Sommer, D. Edwards, Estimating Fatigue from Predetermined Speech Samples Transmitted by Operator Communication Systems, in: Driving Assessment Conference 2009, University of Iowa Libraries Publishing, 2009. URL: https://pubs.lib.uiowa.edu/driving/article/id/28109/. doi:10.17077/drivingassessment.1359.

[71] S. Shinohara, H. Toda, M. Nakamura, Y. Omiya, M. Higuchi, T. Takano, T. Saito, M. Tanichi, S. Boku, S. Mitsuyoshi, M. So, A. Yoshino, S. Tokuno, Evaluation of emotional arousal level and depression severity using voice-derived sound pressure change acceleration, Scientific Reports 11 (2021) 13615. URL: https://www.nature.com/articles/s41598-021-92982-7. doi:10.1038/s41598-021-92982-7.

[72] J. C. Mundt, A. P. Vogel, D. E. Feltner, W. R. Lenderking, Vocal Acoustic Biomarkers of Depression Severity and Treatment Response, Biological Psychiatry 72 (2012) 580–587. URL: https://linkinghub.elsevier.com/retrieve/pii/S0006322312002636. doi:10.1016/j.biopsych.2012.03.015.

[73] J. C. Mundt, P. J. Snyder, M. S. Cannizzaro, K. Chappie, D. S. Geralts, Voice acoustic measures of depression severity and treatment response collected via interactive voice response (IVR) technology, Journal of Neurolinguistics 20 (2007) 50–64. doi:10.1016/j.jneuroling.2006.04.001.

[74] M. Yamamoto, A. Takamiya, K. Sawada, M. Yoshimura, M. Kitazawa, K.-c. Liang, T. Fujita, M. Mimura, T. Kishimoto, Using speech recognition technology to investigate the association between timing-related speech features and depression severity, PLOS ONE 15 (2020) e0238726. URL: https://dx.plos.org/10.1371/journal.pone.0238726. doi:10.1371/journal.pone.0238726.

[75] T. B. Holmlund, A. S. Cohen, J. Cheng, P. W. Foltz, J. Bernstein, E. Rosenfeld, B. Laeng, B. Elvevåg, Using Automated Speech Processing for Repeated Measurements in a Clinical Setting of the Behavioral Variability in the Stroop Task, Brain Sciences 13 (2023) 442. URL: https://www.mdpi.com/2076-3425/13/3/442. doi:10.3390/brainsci13030442.

[76] M. Cannizzaro, B. Harel, N. Reilly, P. Chappell, P. J. Snyder, Voice acoustical measurement of the severity of major depression, Brain and Cognition 56 (2004) 30–35. URL: https://linkinghub.elsevier.com/retrieve/pii/S027826260400140X. doi:10.1016/j.bandc.2004.05.003.

[77] C. A. De Vasconcelos, M. N. Vieira, G. Kecklund, H. C. Yehia, Speech Analysis for Fatigue and Sleepiness Detection of a Pilot, Aerospace Medicine and Human Performance 90 (2019) 415–418. URL: https://asma.kglmeridian.com/view/journals/amhp/90/4/article-p415.xml. doi:10.3357/AMHP.5134.2019.

[78] R. R. Chaturvedi, M. Angrisani, W. M. Troxel, M. Jain, T. Gutsche, E. Ortega, A. Boch, C. Liang, S. Sima, A. Mezlini, E. J. Daza, M. Boodaghidizaji, S.-c. Suen, A. R. Chaturvedi, H. Ghasemkhani, A. M. Ardekani, A. Kapteyn, American Life in Realtime: Benchmark, publicly available person-generated health data for equity in precision health, PNAS Nexus 4 (2025) pgaf295. URL: https://academic.oup.com/pnasnexus/article/doi/10.1093/pnasnexus/pgaf295/8275735. doi:10.1093/

pnasnexus/pgaf295.
[79] J. Haltaufderheide, R. Ranisch, The ethics of ChatGPT in medicine and healthcare: a systematic review on Large Language Models (LLMs), npj Digital Medicine 7 (2024) 183. URL: https://www.nature.com/articles/s41746-024-01157-x. doi:10.1038/s41746-024-01157-x.
[80] E. R. , S. I. , Improving Fatigue Detection with Feature Engineering on Physical Activity Accelerometer Data Using Large Language Models, 2024. URL: https://doi.org/10.60401/ijabc.18. doi:10.60401/ijabc.18.