The Power of Speech in the Wild: Discriminative Power of Daily Voice Diaries in Understanding Auditory Verbal Hallucinations Using Deep Learning

WEICHEN WANG, Dartmouth College, USA
WEIZHE XU, University of Washington, USA
AYESHA CHANDER, University of Washington, USA
SUBIGYA NEPAL, Dartmouth College, USA
BENJAMIN BUCK, University of Washington, USA
SERGUEI PAKHOMOV, University of Minnesota, USA
TREVOR COHEN, University of Washington, USA
DROR BEN-ZEEV, University of Washington, USA
ANDREW CAMPBELL, Dartmouth College, USA

Mobile phone sensing is increasingly being used in clinical research studies to assess a variety of mental health conditions (e.g., depression, psychosis). However, in-the-wild speech analysis – beyond conversation detecting – is a missing component of these mobile sensing platforms and studies. We augment an existing mobile sensing platform with a daily voice diary to assess and predict the severity of auditory verbal hallucinations (i.e., hearing sounds or voices in the absence of any speaker), a condition that affects people with and without psychiatric or neurological diagnoses. We collect 4809 audio diaries from N=384 subjects over a one-month-long study period. We investigate the performance of various deep-learning architectures using different combinations of sensor behavioral streams (e.g., voice, sleep, mobility, phone usage, etc.) and show the discriminative power of solely using audio recordings of speech as well as automatically generated transcripts of the recordings; specifically, our deep learning model achieves a weighted f-1 score of 0.78 solely from daily voice diaries. Our results surprisingly indicate that a simple periodic voice diary combined with deep learning is sufficient enough of a signal to assess complex psychiatric symptoms (e.g., auditory verbal hallucinations) collected from people in the wild as they go about their daily lives.

CCS Concepts: • **Human-centered computing** → **Ubiquitous and mobile computing systems and tools**; • **Applied computing** → **Life and medical sciences**.

Additional Key Words and Phrases: Mobile Sensing, Auditory Verbal Hallucinations, Daily Voice Diaries, Speech in the Wild

Authors' addresses: Weichen Wang, Dartmouth College, Computer Science, Hanover, NH, 03755, USA, weichen.wang.gr@dartmouth.edu; Weizhe Xu, University of Washington, Biomedical Informatics and Medical Education, Seattle, WA, USA; Ayesha Chander, University of Washington, Department of Psychiatry and Behavioral Sciences, Seattle, WA, USA; Subigya Nepal, Dartmouth College, Computer Science, Hanover, NH, 03755, USA; Benjamin Buck, University of Washington, Department of Psychiatry and Behavioral Sciences, Seattle, WA, USA; Serguei Pakhomov, University of Minnesota, , Minneapolis, MN, USA; Trevor Cohen, University of Washington, Biomedical Informatics and Medical Education, Seattle, WA, USA; Dror Ben-Zeev, University of Washington, Department of Psychiatry and Behavioral Sciences, Seattle, WA, USA; Andrew Campbell, Dartmouth College, Computer Science, Hanover, NH, 03755, USA.

2474-9567/2023/9-ART133 $15.00
https://doi.org/10.1145/3610890

ACM Reference Format:
Weichen Wang, Weizhe Xu, Ayesha Chander, Subigya Nepal, Benjamin Buck, Serguei Pakhomov, Trevor Cohen, Dror Ben-Zeev, and Andrew Campbell. 2023. The Power of Speech in the Wild: Discriminative Power of Daily Voice Diaries in Understanding Auditory Verbal Hallucinations Using Deep Learning. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 7, 3, Article 133 (September 2023), 29 pages. https://doi.org/10.1145/3610890

1 INTRODUCTION

Auditory Verbal Hallucinations (AVHs) are a sensory experience ranging from auditory imagery and vivid thoughts to fully developed hallucinations of hearing sounds and voices in the absence of any speaker [29, 58]. Although people with serious mental illness are more likely to experience AVHs (e.g., 60–80% of people diagnosed with schizophrenia [3]), AVH is also experienced by people with a variety of mental health diagnoses (e.g., anxiety, depression, functional impairment [42, 52]) as well as by people who have no identifiable psychiatric or neurological diseases [27, 66, 105]. It has been observed that individuals living with AVH require varying levels of treatment and needs and their clinical status can change throughout the course of their lives [58]. As a result, AVH represents a complex set of experiences that occur in a variety of forms across a wide range of people. Given this, there is a need to better understand the impact and severity of AVH across different population groups from those diagnosed with serious mental illness to those that have no such identifiable psychiatric impairment [29].

Retrospective self-report measures are commonly used in clinical settings to gain insights into the subjective experiences of auditory verbal hallucinations (AVH). However, these measures pose unique challenges. Patients with severe AVH symptoms often come from economically disadvantaged and socially isolated communities that lack adequate treatment resources [74, 99]. Additionally, some patients may be reluctant to seek treatment due to disagreement with the diagnosis or negative attitudes towards mental health services [70]. Those with less severe AVH may choose not to engage in traditional clinical settings due to stigma-related concerns [44, 46]. To address these limitations, recent clinical research projects have turned to the use of mobile phones and wearable technology to gather continuous behavioral data [8]. These studies have demonstrated connections between mobile sensing and behavioral markers of mental health [6, 80, 96, 114].

Despite the advantages offered by existing mobile sensing platforms, such as StudentLife [115], Aware [39], Beiwe [86], and others, in accurately monitoring patients' behavior in everyday life, there has been limited work on analyzing patients' speech collected from mobile phones. The result is that current mobile sensing-based approaches are overlooking valuable information that could be obtained from everyday human speech in the naturalistic environment. Research suggests that speech patterns, such as incoherence [4], limited pitch variation [50], prolonged pause duration [92], and lexical diversity [22, 28, 126], can provide valuable insights into mental health. For example, Minor et al. [78] used an "electronic audio recorder" in real-world settings to better understand psychosis risk, demonstrating the feasibility of using audio recordings to assess real-world expressions of personality and functioning in schizotypy. However, the predictive value of this approach has not yet been determined. This paper addresses the question of whether combining the power of speech in the wild with mobile sensing technology can improve our understanding of AVHs, and how much merit it can bring to the current mobile sensing system.

In this paper, we report on the discriminative power of using speech and content (e.g., transcription) to assess the severity of AVH. We augment the *StudentLife* [115] mobile sensing app to solicit speech via a daily voice diary Ecological Momentary Assessment (EMA). Figure 1 shows the user interface of the daily voice diary. The voice EMA randomly prompts the user four times per day between the hours of 9 am and 9 pm. Participants can also open the app at any time to manually respond to the voice diary survey. The EMA allows participants to record a brief audio diary detailing how they are feeling (see Figure 1). The question that prompts the user is quite open in nature. Furthermore, we tell participants not to overthink how they present their thoughts: "It does not have to be perfect. In fact, we don't want it to be. We want it to be real, and true to you. What would you like to share

with the research team?". In our study, we collected one month of mobile sensing data from N=384 participants. More importantly, we obtained a high volume of voice in the naturalistic environment: a total of 4809 voice diaries were collected, of which 3033 are longer than 30 seconds. While we collect and analyze all StudentLife sensing streams (e.g., phone usage, conversational interaction, mobility, activity, sleep), we focus on the power of human speech to assess and predict the severity of AVH. We use the Hamilton Program for Schizophrenia Voices Questionnaire (HPSVQ) [113] as ground truth that measures the severity of AVH. The HPSVQ is a 9-item questionnaire with a five-point Likert scale rated from 0 to 4. It has excellent test-retest reliability (r = .84) and internal consistency (Cronbach's α = .94). A total HPSVQ score above 26 indicates severe AVH. The HPSVQ scores are dichotomized to derive discrete categories as targets for classification. We design a specialized deep learning architecture to handle multimodal sensing streams (including audio, text and other behavioral sensing data) to predict AVH severity. We do this to best understand the discriminative power of voice. Specifically, we compare and contrast the performance of deep learning models built from (i) only mobile sensing data; (ii) only voice diaries (i.e., audio and transcribed text); and finally (iii) a fusion of mobile sensing data and voice diaries. We also consider both manually and automatically transcribed voice diaries in our modeling and analysis.

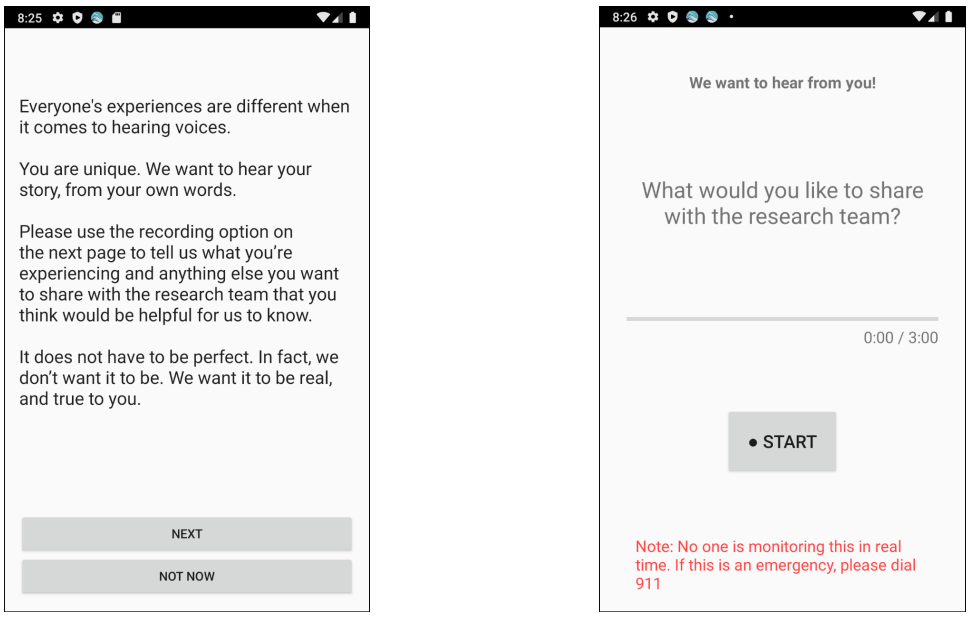


Fig. 1. User interface of the daily voice diary in study app. (A) Welcoming page (B) Recording page.

The contribution of the paper is as follows. First, to the best of our knowledge, we are the first group to study how in the wild daily voice diaries can improve current mobile sensing platforms and predict AVH severity with good performance. We extend an existing and widely used mobile sensing platform for mental health with voice diaries and evaluate various deep neural network architectures step by step, beginning with Bi-direction Gated Recurrent Units (Bi-GRU) [20] and progressing to models with self-attentive embeddings [69], in order to understand the strengths and weaknesses of using each type of data (i.e., sensor data, audio and transcripts) and to find the best architecture that can fully utilize such rich and diverse multi-modal data for assessing AVH severity. Specifically, we propose a deep neural network architecture with self-attention that achieves a weighted f-1 score of 0.84 combining sensor data, audio data of the speech and professionally generated transcripts, and a

macro f-1 score of 0.81 when manually generated transcripts using an automated speech recognition transcription method. A deep neural network model based solely on mobile sensing achieves a weighted f-1 score of 0.65, while a model based solely on voice diary (i.e., audio + text) achieves a weighted f-1 score of 0.78. Our results indicate that periodic voice diaries combined with deep learning alone have sufficient power for assessing a complex condition such as AVH in the wild.

The structure for the rest of the paper is as follows. We describe the related work in Section 2. In Section 3, we discuss our study and data collection. We describe how we preprocess our data in Section 4 and Section 5. We present our deep learning architecture and performance results in Section 6. We investigate the interpretation in Section 7. We discuss the performance of using a fully automatic pipeline using Automated Speech Recognition, as well as the implication of our work in Section 8. Finally, we offer some concluding remarks in Section 8.

2 RELATED WORK

Mobile technologies have the advantage of passively collecting context-based in-situ data, which can be utilized to make various inferences regarding user behavior. The widespread use of smartphones, equipped with a range of embedded sensors such as accelerometers, GPS, microphones, and cameras, provides an opportunity to gain deep and continuous insights into the behavior of individuals diagnosed with mental health illnesses. Clinical research projects have recently embraced mobile phones and wearable sensing technology to collect continuous behavioral data. Studies have found correlations between behavioral sensor data obtained from smartphones and various mental health conditions, including anxiety [14, 96], depression [82, 88, 97, 118, 121], bipolar disorder [47, 94], and serious mental health conditions such as schizophrenia [11, 12, 117, 120]. These findings demonstrate the potential of mobile technologies for facilitating mental health diagnosis and management.

Much of the prior research in the field of mental health and speech primarily operates in controlled laboratory settings, such as recording individuals with schizophrenia reading emotionally neutral text [92]. These studies have demonstrated that linguistic content (e.g., lexical diversity, patterns of word usage, semantic coherence, verbal fluency) [22, 28, 126] and acoustic parameters (e.g., intonation, pitch, jitter, energy [25]) can both provide insights into mental health. Some early attempts have been made to collect speech-related features via mobile sensing. In one of the initial studies, Muaremi et al [83] employed phone call statistics, social signals extracted from phone conversations, and acoustic features of the voice to predict bipolar states in a cohort of 12 individuals who suffered from bipolar disorder. The CenceMe app [77] implemented an always-on voice activity detector and higher-level conversation classifier on the phone to passively determine how much conversation a user is around. Early work by Lu [71] used passive speech analysis to model stress on an Android phone using prosody, pitch, volume, and intonation. Recently, Verily found [85] pairwise correlations between four basic audio parameters (average pause duration, duration of self-recorded audio) and transcribed text (speaking rate, sentiment score) from a 12-week depression research. These works are limited in scope and provide only a limited understanding of speech-related behaviors. To date, there has been no systematic investigation into the full predictive power of speech in natural settings and how it can be leveraged through mobile sensing technology.

Within the field of neural science, there are proponents who assert that auditory verbal hallucinations (AVH) are indicative of an auditory dysfunction [75]. However, the majority of scholars associate this form of hallucination with a speech disorder [43, 53]. It has been discovered that the dysfunction of brain regions responsible for speech production may serve as a fundamental mechanism for the occurrence of AVH in individuals with schizophrenia [107]. Empirical data suggests the presence of electrophysiological abnormalities [68] in the cortices responsible for auditory and speech perception, specifically in Wernicke's area, which is associated with speech perception, and Broca's area, which is associated with speech expression [51, 95]. Studies on auditory perception in individuals with schizophrenia spectrum disorders have revealed impairments in tone matching and pitch discrimination performance for non-verbal sounds [33]. Neuroimaging studies have also found revealed that in individuals with

AVH, neural activation is lateralized towards the language-related areas of the right hemisphere [106]. However, in the majority of individuals, language production is primarily associated with activation in the left hemisphere [65]. The linguistic abilities of individuals diagnosed with schizophrenia have been found to be compromised in several ways, including decreased embedding, shorter phrase length, and increased grammatical errors [41, 81]. Individuals who experience clinical AVH tend to utilize a reduced number of determiners and prepositions, employ shorter utterances, and exhibit a greater prevalence of negative content [23].

The rapid advancement in speech processing technology enables the use of virtual assistants such as Google Assistant, Amazon's Alexa, and Apple's Siri [67]. These virtual assistants rely on deep learning algorithms and always-on speech processors to recognize hot words (e.g., Ok Google) and initiate interaction with the user. However, this method of collection is not practical for mental health studies due to the inability to differentiate between consented and non-consented speech. Instead, we consider Ecological Momentary Assessment (EMA), an alternative approach for collecting real-time data (including human speech) on AVH. Previous studies using EMA have demonstrated correlations between AVH and factors such as time of day [13], anxiety [49], emotional state [30], and cardiac autonomic control [62]. A mobile application that employs EMA to collect AVH-related ground truth from participants has also been developed [104]. Given the potential of vocal and lexical features of speech as predictors of AVH, a useful next step involves examining predictors of AVH using these attributes.

3 STUDY AND DATASET

3.1 Procedures and Participants

The study was approved by the Institutional Review Board (IRB) of the institution [Institution details omitted during anonymous review]. Participants were recruited remotely over the Internet or through community-based strategies. Participants had to be at least 18 years old, English speakers residing in the United States, and have AVH at least once a week [1]. It was also required that they have an Android phone and a data plan. **Online recruiting** was carried out with the help of Google Ads. Google displayed our ads to those whose search history matched pre-defined keywords. We chose keywords based on a study of the literature [24], consultation with researchers on mental illness, and a survey of blogs written by people living with AVH. The keywords contained both clinical terminology (e.g., schizophrenia, bipolar disorder, hearing voices) and non-clinical explanations of AVH (e.g., talking to ghosts, going crazy). If participants clicked on the Google advertisements, they were sent to our recruitment website. The website included detailed infographics and videos that describe the projects in depth. On our website, participants were able to verify their phone number, answer screening questions, agree to the study, complete the baseline evaluation, and download the study Android app. **Community-based recruiting** were through flyers and referrals. Our research staff prescreened the participants over phone and scheduled in-person study visits. During the visit, research staff directed participants to the study's website and assisted them with the same procedures as those who were recruited online.

A total of 384 individuals with AVH (305 recruited online; 79 recruited in the local community) from 41 US states participated in the study and completed data collection. All participants were required to carry a smartphone and respond to EMAs for 30 days. Only if a participant adhered to the 30-day data collection period is he or she regarded to have completed the data collection (those who dropped out during the study are not included in the 384 participants). Participants could directly contact research personnel to ask questions or get technical help. The data were firstly stored locally on the phone and transferred to secure research servers when the phone

[1]The National Institute of Mental Health's Research Domain Criteria (RDoC) [26] is a research framework that guides AVH research on a continuum [40]. Our research, informed by the RDoC framework, focuses on the phenomenology of AVH to elucidate the distinctions between individuals who experience severe AVH symptoms in the context of a clinical disorder and those who encounter AVH without the need for clinical intervention. Therefore, the screening questionnaire is designed to exclude non-AVH individuals, as they fall outside the scope of this study.

detects a connection to the Internet. The study app ceased delivering data to the study team after 30 days. All participants were compensated $125 for participating. Table 1 shows the demographics of the participants.

Table 1. Demographics of the participants.

Category	Count	Percentage
Sex		
Female	192	50.0%
Male	176	45.8%
Transgender: MTF	7	1.8%
Transgender: FTM	5	1.3%
Other:	4	1.0%
Race		
White	238	62.0%
Black or African American	80	20.8%
Asian	7	1.8%
American Indian / Alaska Native	6	1.6%
More than one race	49	12.8%
Missing / Declined	4	1.0%
Ethnicity		
Hispanic / Latino	58	15.1%
Not Hispanic / Latino	324	84.4%
Missing / Declined	2	0.5%

3.2 Ground Truth

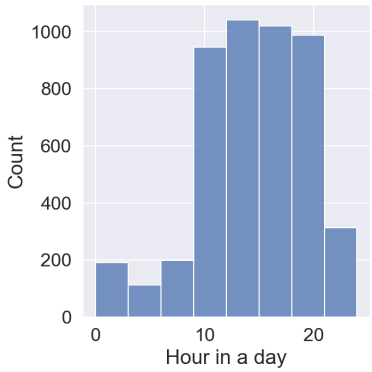
We administered the Hamilton Program for Schizophrenia Voices Questionnaire (HPSVQ) [113] once at the start of the study period (i.e., during the enrollment). HPSVQ is a 9-term self-report measure that received psychometric support [113] for assessing AVH severity. The nine 5-point rating scale components quantify AVH features such as frequency, negative content, loudness, duration, interference with life, distress, influence on self-appraisal, and command compliance. A higher score indicates more severe auditory hallucinations. HPSVQ has shown high test-retest reliability and internal consistency. Furthermore, it has exhibited good concurrent validity [113] when compared to the widely used interviewer-rated Psychotic Symptoms Rating Scales (PSYRATS) [48]. HPSVQ indicates a total score of 0 to 13 for minimal/mild, 14 to 25 for moderate, and 26 and above for severe levels. The ground truth of prediction was derived from these categories.

According to the HPSVQ, 27.4% of individuals in our study exhibited severe AVH (score > 25). Those with scores <= 25, encompassing minimal, mild, or moderate AVH, are classified as having non-severe AVH. Severe AVH is associated with depression, functional impairment, and schizophrenia spectrum disorders (SSDs) [52] and often requires clinical intervention. Identifying factors that distinguish severe AVH holds clinical significance, as it may contribute to the development of targeted treatments for individuals with AVH requiring intervention, as well as preventive strategies for subclinical populations so they do not progress in the psychotic trajectory. Consequently, we opt for a meaningful threshold suggested by the HPSVQ for our severe/non-severe AVH binary classifier's ground truth, rather than employing a threshold from our dataset that yields balanced 50-50 labels.

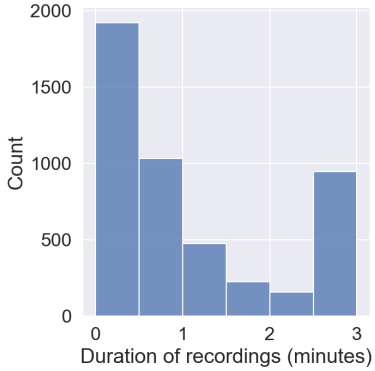
3.3 Dataset

The Android study application is an upgraded version of a system used and validated in prior research [10, 79, 111, 115, 119, 120] that has been tailored to the demands of this study. The study application captures signals

from a variety of embedded sensors (e.g., GPS, screen lock/unlock, light, microphone, accelerometer) to infer participant behavior (e.g., mobility, phone usage, physical activity, etc). We further developed a dashboard for research assistants to monitor user compliance.



(a) Number of voice diaries recorded at different time of the day

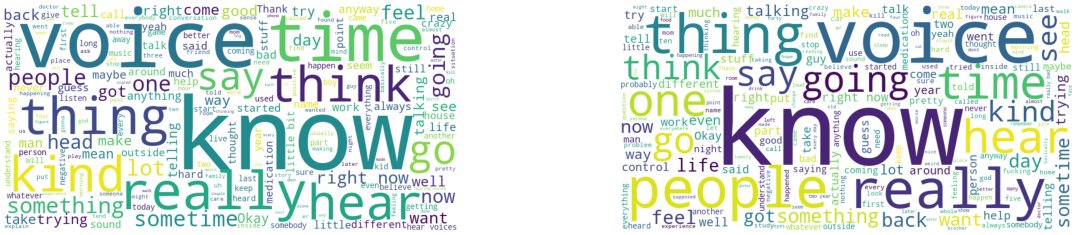


(b) Duration of voice diaries recorded

Fig. 2. We received 4809 audio diaries in all. (a) The majority of the voice diaries were recorded between the hours of 9 a.m. and 9 p.m., when the voice EMA randomly prompts. Others were recorded before or after the typical EMA time. (b) Many voice diaries are less than 30 seconds and contain simply random noise or a few words that are insufficient for analysis. There are also a significant number of audio diaries ranging in length from 2.5 minutes to 3 minutes (the maximum length allowed by the application).

We collected 30-day passive sensing data from 384 participants. We asked participants to optionally record a voice diary to share what they were experiencing or anything else they want to share with the research team. The voice EMA randomly prompts the user four times per day between the hours of 9 am and 9 pm. Participants can also open the app at any time to manually respond to the voice diary survey. Participants can record a voice diary of up to 3 minutes in length. The voice diaries are stored on the phone and uploaded to our secure server backend. Although the voice diary is optional and no extra incentives were offered for it, most participants submitted their recordings. We received a total of 4809 voice diaries, the distribution of which is depicted in Fig.2. There are a number of voice diaries that are less than 30 seconds long and contain only random noise or a few words that are insufficient for analysis and are thus eliminated from the analysis in this study. As such short recordings seldom contain interpretable language, we restrict this collection to recordings lasting 30 seconds or more, leaving 3033 records from 229 users. These 3033 voice diaries were professionally transcribed. The analysis in this paper is based on the 229 users who provided complete 30-day mobile sensing data as well as at least one transcribed voice diary. Among the 229 users, 31.9% exhibited severe AVH, a proportion slightly higher than the 27.4% ratio observed among the total recruited participants, as discussed in Section 3.2. In this paper, we refer to "audio/speech data" as the audio signals from the recordings and "text data" as the transcribed text from the voice diaries. In addition, we used our in-house automatic voice recognition system to create another batch of transcriptions. We also include an analysis in which we compare the results of employing manually and automatically generated transcripts (Section 7.2).

Fig.3 illustrates the word cloud of the audio diaries. Intriguingly, the word cloud plots of people with non-severe and severe AVH are similar; they both notably include keywords such as 'voice' and 'hear,' indicating that we



(a) Audio diaries from people with non-severe AVH

(b) Audio diaries from people with severe AVH

Fig. 3. The word cloud of audio diaries among people with non-severe and severe AVH

need to develop a more comprehensive approach of AVH prediction than only examining the frequency of such words.

4 DATA PREPROCESSING

This data set includes mobile sensing data, audio data from voice diaries, and transcribed text. In this section, we demonstrate how data preprocessing is used to transform unstructured data collected in the wild into well-formed data sets that can be used for further analytics.

Deep learning algorithms, including RNNs, have recently demonstrated exceptional performance without the need for handcrafted features when dealing with sequential data such as text streams, audio snippets, video clips, and time series data [100]. However, because deep machine learning is a black-box technology, the majority of individuals are unable to link the results to an explanation that is human-comprehensible. Such interpretable descriptions are essential for establishing trust between patients and clinicians in the context of practical health.

In our study, we employ both interpretable feature engineering and data preprocessing for deep learning. We perform feature engineering in order to (1) determine the benchmark performance of AVH severity prediction with traditional machine learning models and (2) provide interpretations for representations learned by deep learning models. In addition, we perform data preprocessing for deep learning so that multi-modal time series data can be properly fed to advanced deep models.

4.1 Interpretable Feature Engineering for Traditional Machine Learning

4.1.1 Acoustic Features. We extract acoustic features from speech signals using the open-source toolkit openSMILE (open-source Speech and Music Interpretation by Large-space Extraction) [37, 38]. OpenSMILE is extensively used for automatic emotion recognition in affective computing. We compute acoustic features included in the Geneva Minimalistic Acoustic Parameter Set (GeMAPS) [36], a state-of-the-art standard acoustic parameter set for various areas of automatic voice analysis, such as paralinguistic or clinical speech analysis.

First, 18 low-level descriptors (LLDs) are calculated on each *frame* (a short fragment that contains a few thousands of samples) of audio signals, including:

6 Frequency related parameters
Pitch: logarithmic F0 on a semitone frequency scale, starting at 27.5 Hz.
Jitter: deviations in individual consecutive F0 period lengths.
Formant 1, 2, and 3 frequency: center frequency of first, second, and third formant.
Formant 1: bandwidth of first formant.

3 Energy/Amplitude related parameters
Shimmer: difference of the peak amplitudes of consecutive F0 periods.

Loudness: estimate of perceived signal intensity from an auditory spectrum.
Harmonics-to-noise ratio (HNR): relation of energy in harmonic components to energy in noise-like components.

9 Spectral parameters
Alpha Ratio: ratio of the summed energy from 50-1000 Hz and 1-5 kHz.
Hammarberg Index: ratio of the strongest energy peak in the 0-2 kHz region to the strongest peak in the 2–5 kHz region.
Spectral Slope 0-500 Hz and 500-1500 Hz: linear regression slope of the logarithmic power spectrum within the two given bands.
Formant 1, 2, and 3 relative energy: the ratio of the energy of the spectral harmonic peak at the first, second, and third formant's center frequency to the energy of the spectral peak at F0.
Harmonic difference H1-H2: ratio of energy of the first F0 harmonic (H1) to the energy of the second F0 harmonic (H2).
Harmonic difference H1-A3: ratio of energy of the first F0 harmonic (H1) to the energy of the highest harmonic in the third formant range (A3).

Next, based on the 18 LLDs on each frame from an *utterance*, we generate the 62 acoustic features of high-level statistics functions (HSFs) that will be used in this paper, including:

36 parameters of arithmetic mean and coefficient of variation computed through smoothing all LLDs over time with a symmetric moving average filter 3 frames long.

16 additional functionals based on loudness and pitch over voiced regions representing the 20th, 50th, and 80th percentile, the range of 20th to 80th percentile, and the mean and standard deviation of the slope of rising/falling signal parts.

4 additional parameters over all unvoiced segments: the arithmetic means of the Alpha Ratio, the Hammarberg Index, and the spectral slopes from 0-500 Hz and 500-1500 Hz.

6 temporal features, including the number of loudness peaks per second, the mean length and the standard deviation of continuously voiced regions the mean length and the standard deviation of unvoiced regions (approximating pauses), and the number of continuous voiced regions per second (pseudo syllable rate).

4.1.2 Linguistic Features. Linguistic Inquiry and Word Count (LIWC) [89, 109] is a computerized text analysis software used to quantify language in written or spoken communication. LIWC is commonly used to examine language data and can be applied to the study of language as an indication of psychological, social, and health-related outcomes. We adopt predefined categories in LIWC that have been used in existing pieces of literature that examine mental health from languages [64, 98]. Specifically, we use five categories of, in total 59 linguistic features: (1) **summary language variables**, (2) **basic linguistic style**, (3) **cognitive measures**, (4) **concerns**, (5) **language formality**. **Summary variables** include 4 narrative evaluation: *analytical thinking* [90], *clout* [60], *authenticity* [84], and *emotional tone* [21] and *Word per sentence*. Then, we consider **basic linguistic style** from two aspects: percent of functional words and time orientations. We choose them because the structure of the word and tense are an expression of personality and inner thoughts [91]. Next, we export the **cognitive features**: cognitive processes and perception processes, which directly mirror human perception and thought. In addition, we consider 3 categories of **psychological concerns**: personal concerns, social concerns, and biological concerns as a participant's concerns may reflect their mental states. Lastly, we assess the **formality of language**,

which includes informal speech and all punctuation. These linguistic measures, which are outlined in Table 2, are extracted using the computerized text analysis software LIWC2015 [89, 109].

Table 2. Linguistic features

category		feature
summary language variables		analytical thinking, clout, authenticity, emotional tone, word per sentence (wps)
basic linguistic style	functional words	1st person singular,1st person plural, 2nd person, 3rd person singular, 3rd person plural, impersonal pronouns, articles, prepositions, auxiliary verbs, adverbs, verbs, adjectives, comparisons, interrogatives, numbers
	time orientations	past orientation, present orientation, future orientation
cognitive measures	cognitive process	insight, causation, discrepancy, tentative, certainty, differentiation
	perceptual processes	see, hear, feel
concerns	personal concerns	work, leisure, home, money, religion, death
	social concerns	family, friends, female references, male references
	biological concerns	body, health, sexual, ingestion
language formality	informal speech	swear words, netspeak words
	punctuation	periods, commas, colons, semicolons, question marks, exclamation marks, dashes, quotation marks, apostrophes, parentheses, other irregular marks

4.1.3 Sensor Data Features. We compute features as follows based on raw sensor data:

Physical Activity. We use the Google Activity Recognition Api [45] to determine the activity in which a participant is engaged. This API uses a dynamic algorithm the determine activity using device sensors. We examined 3 variables representing time spent on foot, in a vehicle, and being sedentary.

Phone Usage. Our mobile application tracks the number of phone locks and unlocks performed by participants. We calculate both the total number of phone locks and unlocks and the average time between phone locks and unlocks.

Mobility. The application samples GPS every 10 minutes, taking into account both energy conservation and data quality. Raw GPS coordinates are first clustered using density-based spatial clustering of applications with noise (DBSCAN) [35]. After that, we calculate the number of unique locations and the distance traveled. Furthermore, we compute the time spent in one's primary and secondary locations. Such a strategy has been validated and widely implemented in many research [10, 116] when accessing the lives of people with schizophrenia.

Ambient light. The application measures the ambient light conditions of the user's surrounding environment using light sensors on the phones. This can provide additional contextual information about the environment the user is in.

Sleep. We also infer sleep duration, bedtime, and wake-up time using the method described in [19, 115], which had an accuracy of +/- 32 minutes to the ground truth.

Phone calls and SMS. The application keeps track of SMS text message exchanges (both sent and received), as well as the number and length of phone calls. No written or audio material from these calls or texts was captured to protect privacy. We calculate the number of in/out SMS and calls, and the duration of in/out calls from the logs.

Conversation duration and frequency The study application uses the smartphone microphone to sample and gathers ambient sound in the device's vicinity. A validated [123–125] in-situ privacy-preserving speech classifier determines when human conversations was nearby. To ensure privacy, no raw audio was captured on the device as part of the sensing system, but instead, the data was classified in the time and just the presence of voice was recorded. We compute the number of independent conversations and their duration.

Features above (except for the sleep features) are computed on a daily basis and broken down into four epochs of the day: *morning (6am-12pm)*, *afternoon (12pm-6pm)*, *evening (6pm-12am)* and *night (12am-6am)*, that allow us to model people's behaviors during different parts of the day as recommended by multiple existing research [116, 120]. As a result, we came up with a total of 103 features. Following that, we compute the arithmetic mean of each feature over the 30 days to represent a participant's overall behavior during the research. Days with less than 15 hours of collected data (approximately 9.2% of all days) are considered to have too much missing data and are omitted from the calculation of mean values [117].

4.2 Data Preprocessing for Deep Learning

4.2.1 Acoustic Signals from Voice Diaries. We must first transform the acoustic signal into a series of numbers before we can use it in our models. Our voice diary signals are sampled at 44100Hz. In other words, a single one-second clip would contain 44100 samples, which are too numerous to be processed in their raw state. In fact, any sound frame can be represented by combining signals of various frequencies. And the spectrum - the combination of frequencies that comprise a signal - can be used for further analysis in audio processing using deep learning. Additionally, humans do not perceive frequencies linearly. Variations at lower frequencies are perceived more readily than those at higher frequencies. Similar to frequency, humans perceive loudness logarithmically rather than linearly. Consequently, when dealing with frequencies and amplitudes in our data, we should utilize a logarithmic scale by means of the Mel Scale and the Decibel Scale. Mel Spectrogram is the outcome of these two scale adjustments. We compute the Mel Spectrogram using Librosa, a python-based audio signal processing library. Fig 4 shows the Mel Spectrogram computed from one of our voice diaries. It plots frequency along the y-axis and time along the x-axis.

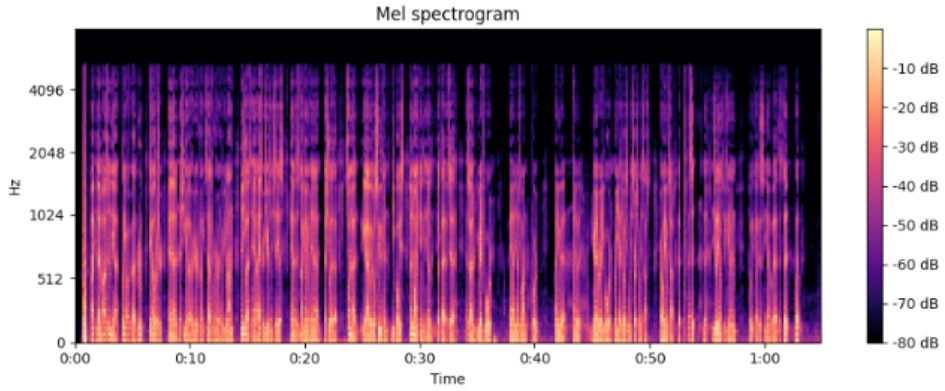


Fig. 4. Mel Spectrogram computed from one of our voice diaries.

We define a frame as a sequence of 8192 values, which corresponds approximately to a 200ms segment in a 44100Hz diary clip. The maximum length of a participant's diary is 3 minutes, resulting in a maximum length T of 180s * 44100Hz / 8192 = 968 in our computed output. We compute 128 Mel scales for each frame (the default

setting of Librosa, which is commonly used in audio signal processing). As a result, the entire output (which will be used as the input for deep models) is a (T = 968) (M = 128) matrix, and the colors in Fig. 4 represent the values. The outputs of diaries with a duration of less than 3 minutes are zero-padded so that the padded values can be identified and bypassed in RNN.

4.2.2 Transcripts from Voice Diaries. Before we can feed the transcripts into our models, we must first convert them to a set of vectors. We use the pre-trained BERT (Bidirectional Encoder Representations from Transformers) model [32] to extract contextual embeddings from the text. BERT is trained on the Wikipedia corpus in order to solve different natural language processing tasks (e.g., named entity recognition). We first tokenize the words in the sentences based on BERT's vocabulary. We then extract a contextualized embedding for each of the tokens. BERT automatically splits the original word into smaller subwords and characters (i.e., tokens). If the input has multiple sentences, BERT uses the special token [SEP] to differentiate them. It also inserts a [CLS] token at the start of the text to indicate the beginning. Each token is then embedded into a 768-dimensional space producing a single 768-length vector. As a result, for each transcript text file, we create a sequence of length T, where T is the number of tokens separated by BERT, and each specific position t in a such sequence is a vector of length 768. All created sequences are 0 padded afterward to match the length of the longest text in the data set, as we must keep the output the same shape for each transcript to feed them into the deep models.

4.2.3 Sensor Data from Smart Phones. Mobile sensing data is commonly used as sequences in two ways: (1) low-level raw sensor signals sampled every second, and (2) high-level aggregated/inferred behavioral variables on an epoch/daily basis. The first method is typically utilized for real-time applications like activity recognition. The second is usually applied in long-term research. We use the second strategy in this study and create a time sequence with a length of T=30 (days). In such a sequence, each position t is a vector containing 103 features generated in Section 4.1.3. In deep learning, we keep the entire sequence instead of computing the arithmetic mean of each feature throughout the 30 days so that the deep model can learn knowledge from the entire time series. Deep neural networks frequently employ normalization procedures to boost convergence and generalization. As a result, for each feature, we compute two normalized (within-person and between-person) values to capture both within-person and between-person differences during the study. Finally, the dimension of the sensor data input is T * M, where T = 30 denotes the number of days in the research and M = 2 * 103 = 206 denotes the number of normalized features computed from sensor signals. All missing values are replaced with -999 so that they can be recognized and skipped in deep recurrent neural networks.

In the following sections of this paper, we refer to the **pre-processed** acoustic signals, transcripted text, and sensor data that we plan to use as inputs to our deep models as "audio data," "text data," and "sensor data" for simplicity.

5 PREDICTING MODELS AND RESULTS

In this section, we evaluate various deep neural network architectures, from simple to complex, in order to comprehend the predictive power of each type of data (sensor data, audio data, and text) and to identify the best architecture capable of utilizing such rich and diverse multi-modal data for assessing AVH severity. The binary labels we used for training and tried to predict are based on the HPSVQ participants completed during the enrollment, as described in Section 3.2. That says, the prediction is made once per participant over the whole duration of the study, which is 30 days. [2]

[2]Identify severe AVH using the 1-month dataset holds diagnostic value. Per the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5) [5], the co-occurrence of AVH symptoms during a 1-month period, along with another "characteristic sympto" such as delusions or disorganized speech, is indicative of a diagnosis of schizophrenia. Furthermore, it should be noted that the HPSVQ is not intended to measure short-term changes, as evidenced by its consistently high test-retest reliability when administered on a weekly basis [61]. In light

To evaluate the models, we selected 60%, 20% and 20% of the independent users, respectively, as the training, validation, and test sets. Data from a single participant appear in only one of the sets (leave-subjects-out). We are aware that each user has multiple sequences of processed audio data and transcripts (depending on the number of voice diaries uploaded), but only one sequence of 30-day sensor data (from Section 3). We are not using all of the audio data and transcripts to train the model because (1) some participants may upload significantly more voice diaries than others, resulting in an unbalanced training set (with a large amount of data from a single participant), and (2) short voice diaries may not contain sufficient information for data analysis. Therefore, we select the 10 voice diaries with the longest audio length for each participant, and then select the 5 with the highest word count in the transcript. This strategy may not be ideal, but it should be satisfactory for choosing a set of most informative voice diaries from each participant, given that we do not intend to infer the occurrence of AVH from individual audio diaries. [3] At each training epoch for users in the training set, one of the user's selected voice diaries is chosen to train the model. For each user in the validating or testing set, we simply select a voice diary at random from those that have been selected. The model is trained with the Adam optimizer [63] and tuned for optimal performance on the validation set. The testing set performance is reported.

5.1 Baseline: Using Traditional Machine Learning Approaches

Before moving on to deep learning approaches (which typically outperform regular machine learning methods on sequential data), we would like to establish a baseline with classic machine learning methods. We train an XGBoost [17] classifier with all of the interpretable features specified in Section 4.1. XGBoost is a highly efficient and flexible gradient-boosting library. We conduct a grid search on specified hyper-parameters to optimize the performance of the model formed from the training set on the validation set. Using the selected super-parameters, we next train a model on the combined training and validation data and evaluate it on the testing set. Our XGBoost model achieves a macro f1-score of 0.68 and a weighted f1-score of 0.72. We will further discuss the important features in Section 6.

5.2 Uni-modal Data

In this subsection, we use the pre-processed audio data, transcripts, and sensor data from Section 4.2 as the input for deep learning models to evaluate the performance of predicting AVH severity from each data type. A masking layer [73] is used immediately after the input to inform the model that certain timesteps in the input are missing (e.g., the 0 padding) and should therefore be skipped when processing the data.

5.2.1 Bidirectional Gated Recurrent Unit (BiGRU). We first predict the AVH using a Bidirectional Gated Recurrent Unit (BiGRU) network [20]. If there are H hidden units in the GRU, the bidirectional GRU would output 2 * H values. The model is constructed as detailed in Fig.5. The orange rectangle represents a vector of length M at a particular time for a sequence shaped T * M. The bidirectional GRU layer produces an output (the red rectangle) with the shape (T, 2 * H), where H is the number of hidden units in the GRU. Thus, we create a simple vector representation of the sequence data. The representation generated by the bidirectional GRU layer is then fed into fully connected neural networks and dropout layers.

---

of the characteristics of AVH, researchers scholars commonly assess the changes in the phenomenological characteristics of AVH through longer-term follow-up, such as 6 months [61] or even up to 1 year [15].

[3]It is assumed that the lengthy voice diaries selected during the month are the most representative for inferring the severity of AVH, while shorter ones may introduce noise to the prediction.

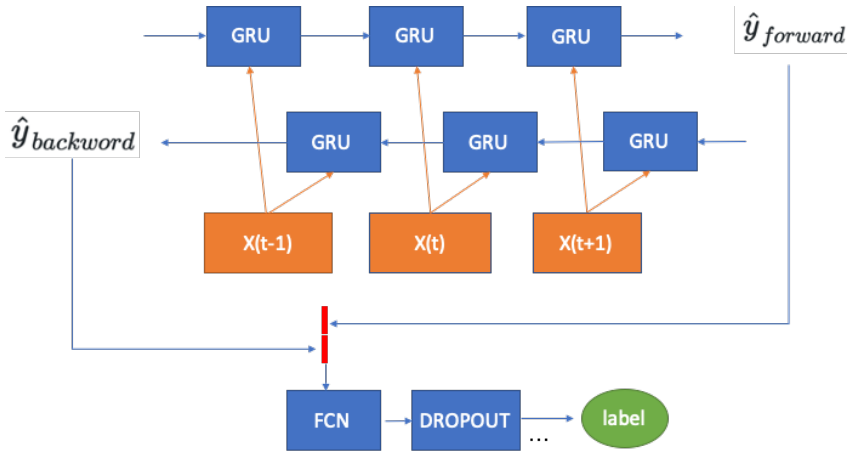


Fig. 5. Illustration of Bidirectional Gated Recurrent Unit network.

5.2.2 Self-attentive Embedding. We augment the bidirectional GRU models with self-attention embedding [69], a widely employed technique for sentence embedding in natural language processing. This attention mechanism enables a recurrent model to focus on distinct portions of a lengthy input stream. The model is constructed in accordance with Fig.6. We anticipate that self-attentive embedding will improve performance when text data is used as input because it has been utilized successfully in text processing. We have yet to determine whether or not it improves performance when audio or sensor data is used as input.

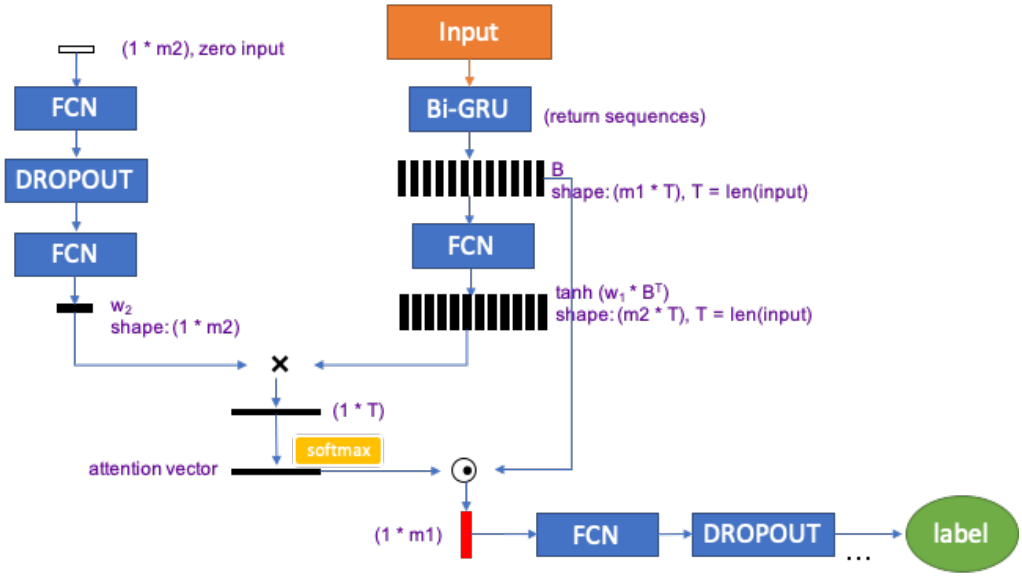


Fig. 6. Illustration of self-attentive BiGRU network.

Instead of relying solely on the final hidden state of the RNN, self-attentive embedding necessitates all hidden states at each step of the RNN. Assuming that the hidden unit number for each unidirectional GRU is H, we will obtain an output of $B = (h_1, h_2, ...h_T)$ from BiGRU, where h_t represents the concatenation of the hidden states from the forward and backward GRU at position t. The attention mechanism encodes B into an embedding by choosing a linear combination of the T BiGRU hidden vectors in B. The linear combination weights are defined in the attention vector α, which is computed as $\alpha = softmax(w_2 * tanh(w_1 * B^T))$. The w_1 and w_2 are learned from the data themselves. Note that w_2 is trained with an initial input of zero because we have no prior knowledge of the attention vector. The softmax ensures that the sum of all computed weights is 1. Then, weighted sums are computed by taking the inner product of the attention vectors a and B. The representation produced by self-attentive BiGRU is then used as the input for subsequent fully connected neural networks and dropout layers to predict the severity of AVH.

5.2.3 Predicting AVH using Uni-modal Data. Table 3 displays the f-1 scores when predicting the severity of AVH using uni-modal data (sensor, audio, or text) and various deep neural networks. As anticipated, our experiments demonstrate that self-attentive models, which were originally utilized in NLP areas, dramatically improve the performance of text data-based models. The text-based self-attentive model achieves a close to 0.7 weighted f-1 score. It marginally improves the audio data a little bit (in terms of macro f-1). This indicates that despite the fact that the mel spectrogram of raw audio data still contains a great deal of prosody-semantic information, it is more difficult to learn a good attentive embedding from the spectrogram than from the transcribed text. In terms of sensor data, the self-attentive model performs even worse than the BiGRU model. This suggests that focusing on different areas of the 30-day study period will not improve the AVH prediction, which is to be expected given the nature of rolling enrollment. Instead, we could simply apply RNN without attentive embedding.

Table 3. Predictive performance on AVH using uni-model data.

f1-score	bigru-sensor	bigru-audio	bigru-text	self-att-sensor	self-att-audio	self-att-text
macro	0.59	0.53	0.52	0.46	0.56	0.68
weighted	0.65	0.63	0.6	0.58	0.58	0.7

5.3 Multi-modal Data

In this subsection, we combine the optimal deep learning architecture for each type of uni-modal data described in Section 5.2 by concatenating the representations from the sensor, audio, and text neural networks prior to feeding them into the final fully connected neural networks and dropout layers (Fig. 7a). In addition, we hypothesize that para-linguistic features (e.g., emotions, speed, energy, tone, etc.), which may be involved in the representation of audio (orange rectangle in Fig. 7b, may influence the attentive embedding in text. Consequently, in Fig. 7b, we use the representation of the audio as prior knowledge about the attention vector for text and learn w_2 (one of the parameters for generating the attention vector, see Fig.6) starting from this representation rather than a zero input.

Table 4 depicts the F-1 scores for predicting the severity of AVH using multi-modal data and the two architectures depicted in Fig. 7. Architecture (b) that uses the audio sequence representation as a starting point for learning the text attentive embedding parameter outperforms architecture (a). The audio + text model achieves a weighted f-1 score of 0.78, indicating the discriminatory power of using the wild voice diaries to determine the severity of AVH. In addition, the addition of passive sensor data from mobile devices eventually results in a better weighted f-1 score of 0.84. Table 5 and Table 6 compare the confusion matrices between the "audio+text" model and the "combine-all" model, using the architecture (b). Based on the analysis of the confusion matrices, it appears that the

inclusion of mobile sensing data has the potential to enhance precision, but does not appear to have a significant impact on recall within our dataset. In our best model, the false positive rate, which refers to the probability of an individual being erroneously diagnosed with severe AVH, is minimal, specifically 6%.

Table 4. Predictive performance on AVH using multi-model data.

f1-score	audio+text (a)	combined all (a)	audio+text (b)	combined all (b)
macro	0.73	0.77	0.75	0.81
weighted	0.75	0.81	0.78	0.84

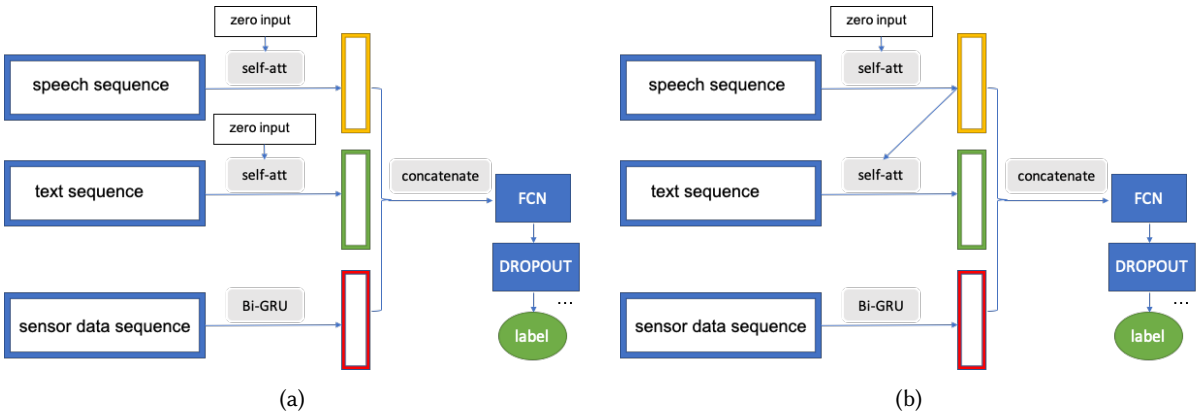


Fig. 7. Illustration of deep neural network using multi-modal data (audio, text and sensor data).

Table 5. Confusion matrix: using architecture (Fig. 7b) and the "audio + text" model (weighted f-1 score = 0.78)

		Prediction 0	Prediction 1	total
Ground-truth	0	*True Neg.* **26**	*False Pos.* **5**	31
	1	*False Neg.* **5**	*True Pos.* **10**	15
	total	31	15	

Table 6. Confusion matrix: using architecture (Fig. 7b) and the "combined-all" model (weighted f-1 score = 0.84)

		Prediction 0	Prediction 1	total
Ground-truth	0	*True Neg.* **29**	*False Pos.* **2**	31
	1	*False Neg.* **5**	*True Pos.* **10**	15
	total	34	12	

6 INTERPRETATION

In this section, we study the feature interpretation for AVH prediction in two ways: the importance of handcrafted features in the XGBoost classifier and the interpretation of deep learning-learned features.

6.1 Important Interpretable Features

SHAP (SHapley Additive exPlanations) is a game-theoretic way to explaining the output of any machine learning model [72]. SHAP is one of the most used post-hoc explainability methods for computing feature attributions. To understand how a single feature affects the AVH prediction, we plot the SHAP values of top-15 important features for every sample in the dataset in Fig 8. The y-axis, from top to bottom, ranks the features in order of importance. The x-axis refers to the actual SHAP values. The horizontal location of a point shows the feature's impact on the model's prediction for a given sample ($x > 0$: more likely to be severe AVH; $x < 0$: more likely to be non-severe AVH). The color of each point represents the value of the feature, with red indicating a high value and blue indicating a low value. This visualization provides a comprehensive overview of the contributions of each feature to the AVH prediction.

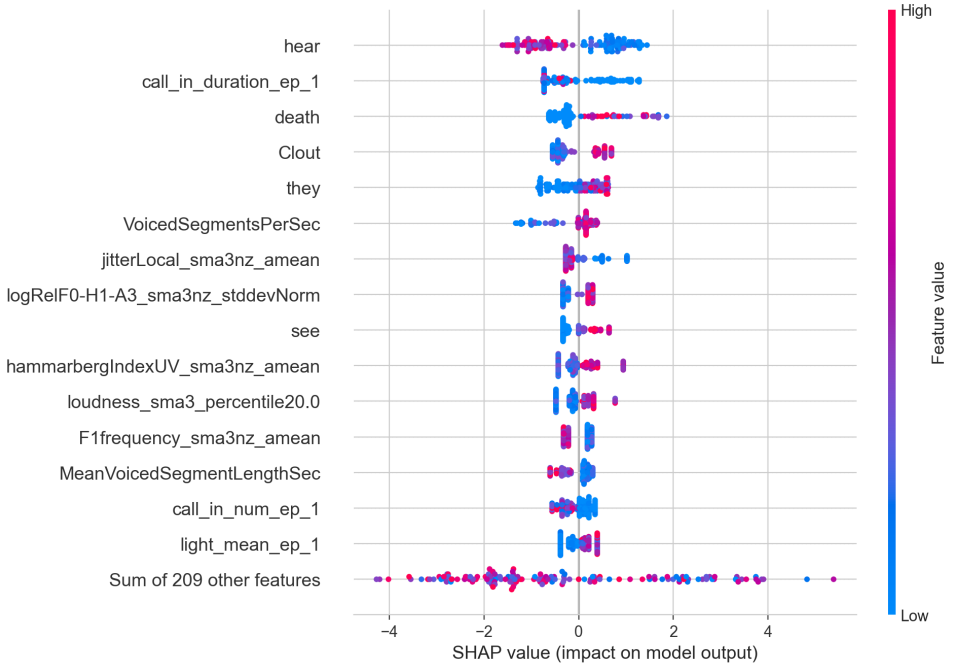


Fig. 8. SHAP value summary plot of the XGBoost model trained on handcrafted features

From the summary plot presented in Fig. 8, it can be observed that the following features have a significant impact on the prediction of a person with severe AVH:
[Audio diary text features] *lower* occurrence of words related to "hearing things"; *higher* occurrence of words related to "death", 3rd person plural, and "seeing things"; having *higher* clout (power in talking).
[Audio diary acoustic features] *larger* number of continuous voiced regions per second, *smaller* mean of the deviations in individual consecutive F0 period lengths, *higher* SD on the ratio of energy of the first F0 harmonic

(H1) to the energy of the highest harmonic in the third formant range (A3), *higher* mean Hammarberg index (the ratio of the strongest energy peaks in the 0–2 kHz vs 2–5 kHz regions) of unvoiced regions, *higher* loudness, *lower* first formant (F1) frequency, *shorter* length of continuously voiced regions.
[Mobile sensing features] *higher* deviation in illuminance during night; *fewer* call-ins during night.

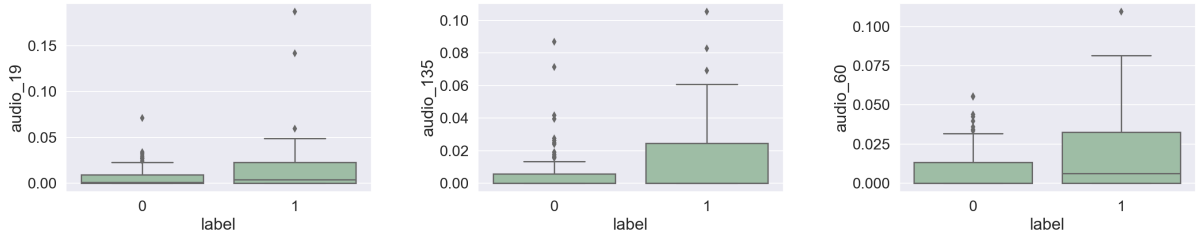
The results of the study provide new insights into AVH, some of which align with established clinical knowledge and hypotheses. For example, words pertaining to "death" may serve as an indicator of the presence of major depressive disorder [9]. "Seeing things" may indicate the presence of visual hallucinations, which typically co-occur in association with other hallucinations such as AVH [122]. The larger number of continuous voiced regions per second and shorter length of continuously voiced regions are consistent with the tendency of clinical AVH patients to employ shorter utterances [23]. Other acoustic features may suggest basic impairments in auditory processing [112] or voice-related distress [102]. The deviation in illuminance during the night and fewer call-ins could indicate social withdrawal [54]. All of the aforementioned factors are associated with AVH. However, some of the results may appear counter-intuitive, such as the lower occurrence of words related to "hearing things" among patients with severe AVH. This may be due to the fact that patients with more severe symptoms often have lower insight, and therefore do not describe their experiences within a clear frame of "hearing things". Some may still be lacking in literature support (e.g., higher clout). Despite these limitations, the important features identified in this study demonstrate the feasibility of using multi-modal data for assessing the severity of AVH. However, it is important to note that the feature importance is based on a baseline model and further validation is needed through additional research to avoid false discoveries. We plan to delve deeper into interpreting the high-level representations automatically learned through deep learning.

6.2 Interpretation of Auto-Learned Features from Deep Learning

It is crucial that we open the "black box" of a deep neural network and explain why it makes decisions in health and wellness applications in order to enhance patient and clinician trust. In this part, we propose a post hoc analysis for clinical interpretation of the learned representations from voice, transcript, and sensor data obtained from our best model (reference Fig. 7b). We export the values in the **hidden units** in the yellow layer (256 hidden units that represent audio input), green layer (256 hidden units that represent text input), red layer (256 hidden units that represent sensor data input), and the fully connected layer (256 hidden-units that represent multi-modal data input) before the last dropout layer. We label each value in the learned representations with the type of data source and its index in the representations; for instance, audio-i denotes the i-th value in the representation learned from audio sequence (yellow rectangle in Fig. 7b) and text-j denotes the j-th value in the representation learned from text sequence (green rectangle in Fig. 7b). We employ ANOVA to examine the **auto-learned features** across groups with non-severe AVH and severe AVH. Based on the effect size (partial eta squared) from ANOVA, the most discriminative auto-learned features of each category of data are chosen. Then, we analyze the relationship between these selected auto-learned features and interpretable hand-crafted features generated in Section 4.1.

6.2.1 Representations Learned from Audio Data. Fig 9 shows the distribution of the top-3 most discriminative auto-learned features from audio data (represented by the green rectangle in Fig. 7b) among the non-severe (label 0) and severe (label 1) AVH groups. People with severe AVH are more likely to have a **higher** value in audio-19, audio-135 and audio-60. The effect sizes are all around 0.06, indicating a medium effect.

The table 7 lists the top-3 hand-crafted features with the strongest correlations (i.e., having the highest absolute values of correlations) to the selected auto-learned audio features. For a detailed explanation of each hand-crafted feature, refer to Section 4.1.1. Our results suggest that the network has learned multiple representations associated with lower pitch and loudness, as well as steeper slopes of both rising and falling portions of loudness. These representations may indicate negative symptoms or distress, respectively, and correspond to severe AVH.



(a) audio-19, $\eta^2 = 0.07$ (b) audio-135, $\eta^2 = 0.06$ (c) audio-60, $\eta^2 = 0.06$

Fig. 9. The distribution of the most discriminative representations from audio data among the non-severe (label 0) and severe (label 1) AVH group. The effect size is indicated by partial eta squared.

Table 7. Top-3 representations automatically learned from audio data of voice diaries that identify patients with severe AVH, as well as the top-3 handcrafted acoustic features with the strongest correlation to the representations.

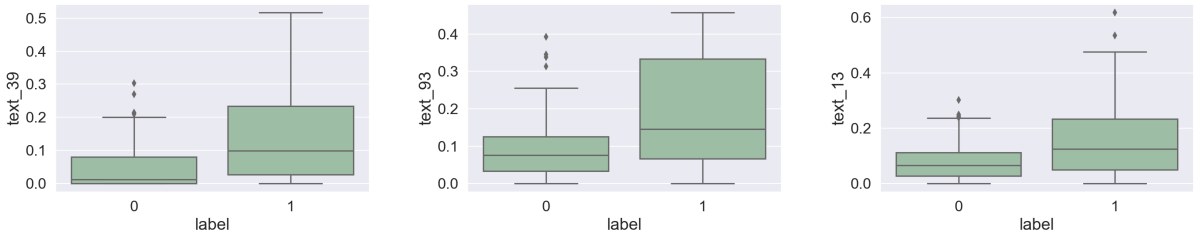
Representations	feature_1	feature_2	feature_3
audio-19	(+0.26) Mean Hammarberg index over voiced regions	(-0.26) Mean Spectral Slope 0-500 Hz over voiced regions	(-0.24) Mean Pitch
audio-135	(-0.26) Mean Spectral Slope 0-500 Hz over voiced regions	(+0.23) Standard deviation of the slope of rising signal parts of Loudness	(+0.22) Mean slope of rising signal parts of Loudness
audio-60	(+0.23) Standard deviation of the slope of falling signal parts of Pitch	(+0.22) Mean slope of falling signal parts of Pitch	(-0.18) Mean Loudness

These findings align with previous research in psychology and neuroscience. For instance, individuals with schizophrenia spectrum disorders often experience deficits in auditory perception, such as pitch-based tone matching [108]. Additionally, these individuals may have event-related potential deficits and neuroanatomical abnormalities in the auditory cortex [56]. Such deficits in basic auditory processing and sensation can impact higher-level cognition and have been linked to hallucinations [55]. Studies have also shown that individuals with auditory hallucinations perform worse on pitch discrimination tasks compared to those without AVH [76].

6.2.2 Representations Learned from Text Data. Fig. 10 illustrates the distribution of the top 3 most discriminative auto-learned features from text data (represented by the green rectangle in Fig. 7b) between non-severe (label 0) and severe (label 1) AVH groups. Individuals with severe AVH tend to exhibit **higher** values for features text-39, text-93, and text-13. The effect sizes suggest a large or close-to-large effect. The network has learned high-level representations that are associated with text data that contains more past tense, fewer present tense expressions, a higher emphasis on personal concerns related to money and work, and fewer references to cognitive and perceptual processes, which may indicate the presence of severe AVH (refer to Table 8).

The focus on personal concerns related to money and work may suggest stressors that could exacerbate symptoms [103]. There have been studies on the relationship between past focus and mental illness such as schizophrenia and mood disorder, and have led to mixed results, with some research showing an increased use [31] and others a decreased use [7] of past tense in written materials focusing on schizophrenia. It's worth noting that these studies relied on written materials, such as online essays and social media content, while our analysis is based on transcripts from oral audio diaries. The findings regarding a lower emphasis on cognitive and perceptual

processes are consistent with the argument by Javitt [55] that sensory deficits can negatively impact cognitive processes.



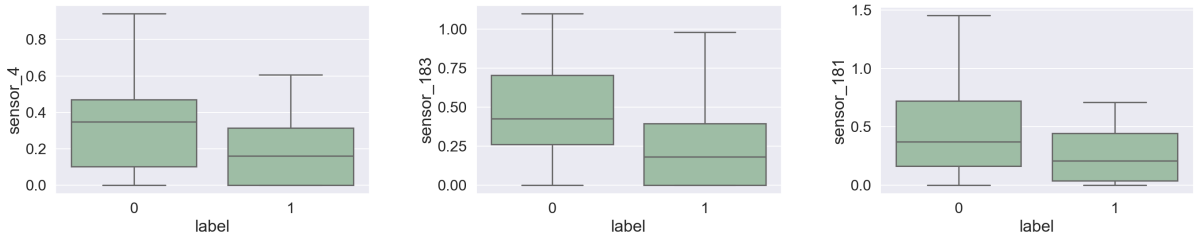
(a) text-39, $\eta^2 = 0.16$ (b) text-93, $\eta^2 = 0.13$ (c) text-13, $\eta^2 = 0.13$

Fig. 10. The distribution of the most discriminative transcript representations between the non-severe (label 0) and severe (label 1) AVH groups, effect size indicated by partial eta squared.

Table 8. Top-3 representations automatically learned from text data from voice diaries that differentiate patients with severe AVH the most effectively, as well as the top-3 created linguistic features with the strongest correlation to the representations.

Representations	feature_1	feature_2	feature_3
text-39	(+0.37) Past tense	(+0.30) Personal concerns: Money	(-0.29) Present tense
text-93	(-0.31) Perceptual processes: Hear	(+0.27) Personal concerns: Money	(-0.25) Perceptual processes: All
text-13	(+0.27) Personal concerns: Work	(+0.27) Personal concerns: Money	(-0.27) Perceptual processes: All

6.2.3 Representations Learned from Mobile Sensing Data. Fig 11 shows the distribution of the top-3 representations auto-learned representations from mobile sensing data (red rectangle in Fig. 7b), obtained from smartphones, that effectively distinguish patients with severe AVH. Patients with severe AVH tend to have **lower** values for sensor-4, sensor-183, and sensor-181. The effect sizes, as measured by partial eta squared, suggest a medium effect.



(a) sensor-4, $\eta^2 = 0.08$ (b) sensor-183, $\eta^2 = 0.08$ (c) sensor-181, $\eta^2 = 0.07$

Fig. 11. The distribution of the most discriminative sensing data representations between the non-severe (label 0) and severe (label 1) AVH groups.

Table 9 lists the top-3 hand-crafted features with the strongest correlation to the auto-learned representations from mobile sensing data. It is important to note that the coefficient listed in the table represents the association between the hand-crafted features and the auto-learned representations. Given that severe AVH is correlated with lower values of the representations (Figure 11), readers should consider the inverse coefficients when examining the relationships between the hand-crafted features and severe AVH.

Table 9. Top-3 hand-crafted features with the strongest correlation to the auto-learned representations from mobile sensing data.

Representations	feature_1	feature_2	feature_3
sensor-4	(-0.27) Duration of conversations nearby (night, 0am-6am)	(-0.27) Sleep: wake up time	(-0.27) Number of conversations nearby (0am-6am)
sensor-183	(+0.48) Duration of out-going phone calls (whole day)	(+0.45) Duration of out-going phone calls (morning, 6am-12pm)	(+0.45) Number of in-coming phone calls (whole day)
sensor-181	(+0.35) Duration of phone usage (evening, 6pm-0am)	(+0.32) Duration of in-coming phone calls (whole day)	(+0.32) Duration of phone usage (whole day)

Sensor-4 represents night-time conversations and wake-up time, both of which may indicate insomnia, a known association with auditory hallucinations [101]. Sensor-183 pertains to telephone calls, while Sensor-181 focuses on mobile device usage. In general, the findings suggest that sleeping in and waking up late, reduced incoming and outgoing phone calls, and decreased phone usage could indicate severe AVH, potentially reflecting social isolation [54], which has been proposed to have a mechanistic relationship with psychotic symptoms.

6.3 Learned Attention Vector in Sentence Embedding

The use of self-attention in text enables an interpretation of the learned sentence embedding: we can generate heat maps of the weight vectors associated with the text, allowing us to visualize which portions of the audio diary are considered when predicting AVH. In Figure 12, we present the attention vector for a selected segment of the audio diary. The darker areas on the heat map indicate the clauses or phrases that carry more weight in the attention vector. The model appears to capture significant factors in the audio diary that are related to the severity of AVH. These factors include *following voices* ("maybe it's that I get control"), *interfering voices* ("but the thing is that it starts to come on strong"), *loud/distressing voices* (feeling like you want a gallon scream at the top), and *feeling bad of self* ("I'm so frustrated, it's a very confusing illness"), all of which are encompassed in HPSVQ.

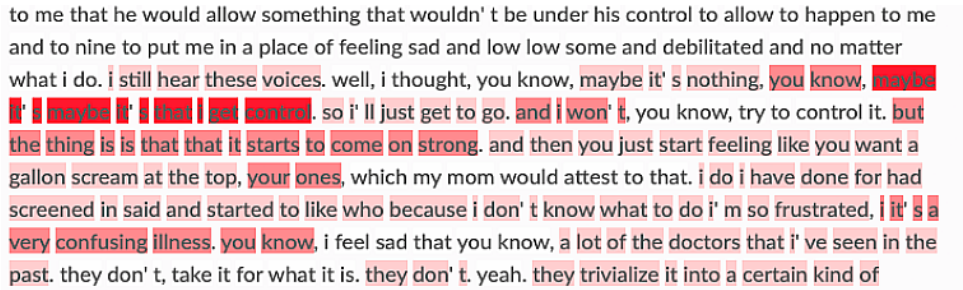


Fig. 12. Attention vector of a segment of an audio diary.

7 DISCUSSION

7.1 Using Multimodal Streams and Deep Learning to Predict AVH

The purpose of our research was to identify the best deep learning architecture for predicting AVH using multimodal stream data. To accomplish this, we conducted experiments using various deep learning approaches.

In contrast to traditional machine learning algorithms, deep learning models have the advantage of being able to handle time series data more effectively. Traditional algorithms require manual extraction of a large number of statistical features from each time series, leading to loss of information and a high number of correlated features. Moreover, there is no way to assign different time periods in the series different weights, because it is difficult to determine which time periods are most relevant for predicting outcomes. Our experiments showed that deep learning models outperformed XGBoost, confirming their suitability for this type of problem.

Traditional machine learning algorithms may struggle with time series data. Researchers must compute multiple hand-crafted or aggregated statistical features on each time series data (e.g., average, percentile, maximum, minimum, etc.), resulting in information loss and an explosion of highly correlated features. In addition, there is no way to assign different periods in the time series different weights because we do not know which time period is most relevant for predicting a particular outcome. Our experiment also shows that deep learning outperforms the XGBoost.

The results presented in Table 3 demonstrate that the text-based model is more effective compared to those relying on speech and sensor data. This highlights the significance of participant's words (i.e. content) as better indicators of AVH, compared to the acoustic parameters in speech and daily behavior recorded by mobile phone sensors. This finding is noteworthy, particularly in light of previous studies that have predominantly focused on predicting symptoms of schizophrenia through passively sensed behavior [1, 111, 117]. However, predicting the severity of AVH can prove to be a *more complex* task, as it is a symptom that can occur not just in individuals with serious psychotic disorders like schizophrenia, but also in healthy individuals. In fact, as many as 75% of individuals who do not meet the criteria for a psychotic disorder have reported experiencing AVH [59]. Moreover, although AVH is a hallmark symptom of schizophrenia, not all patients with schizophrenia experience AVH [18]. Some neuroscience studies have shown that compared to schizophrenia patients without AVH (non-AVH), those with AVH exhibit distinct intrinsic connectivity patterns in cortico-subcortical circuits [93] and interhemispheric circuits [16]. However, the underlying mechanisms by which AVH arise spontaneously from intrinsic brain activity remain elusive. Our results imply that predicting AVH based solely on passively sensed behavior may not be sufficient, and demonstrate the efficacy of utilizing voice diary sequences.

7.2 Fully Automatic Pipeline (Using Automated Speech Recognition)

So far, in our investigations, we have utilized text that has been manually transcribed. This method enables us to grasp the potential of voice diary transcriptions. Nevertheless, manual transcription services are not feasible due to their high cost and manual nature. Thus, we are now further exploring the use of Automated Speech Recognition (ASR) in a fully automated system.

With the aim of preserving patient confidentiality, we have chosen to utilize an in-house ASR system instead of relying on audio-to-text services from third-party providers. This in-house system [127] is based on Baidu's Deep Speech 2 architecture [2] and was implemented using PyTorch [87]. The system was trained with speech corpora obtained from the Linguistic Data Consortium [4], a consortium that includes universities, libraries, corporations, and government research laboratories.

The results of our AVH prediction performance using a fully automated system with ASR instead of manual transcription services can be seen in Table 10. Although it is expected that the accuracy of the prediction may

[4]https://www.ldc.upenn.edu/

decrease due to errors in the ASR transcription process, our results show that even with sensor data, audio, and transcripts obtained through this automated approach, we still obtained a macro f-1 score of 0.76 and a weighted f-1 score of 0.80. These findings suggest that it is possible for the entire AVH prediction process to be carried out with minimal human involvement in the future.

Table 10. Predictive performance on AVH using multi-model data (with text data from ASR). See Fig. 7 for the definition of structure (a) and (b).

f1-sccore	audio+text (a)	combined all (a)	audio+text (b)	combined all (b)
macro	0.64	0.76	0.72	0.76
weighted	0.71	0.80	0.78	0.80

7.3 Implications

Despite the fact that researchers in the ubiquitous computing community (many of whom partnered with clinical researchers) have investigated mobile sensing-based technology for a number of years [8], it has not yet been widely adopted in the clinical treatment process. This paper presents an alternative method to enhance existing mobile sensing-based applications by utilizing a simple periodic voice diary to assess the severity of AVH in the wild.

AVH is the perceptual experience of hearing sounds or voices. This type of auditory hallucination may cause considerable distress and disability. 5-28% of the general population experiences AVH [29, 57, 110], However, it is challenging for medical professionals to assess the severity of AVH. Not all patients wish to or are able to contact physicians due to a variety of factors (e.g., stigma, discrimination. lack of access to services). Fewer than 40% of people with schizophrenia receive treatment from mental health professionals [34], let alone those with AVH, which can be experienced by individuals with no diagnosed psychiatric disorders. Theoretically, the proposed automated systems can be reached remotely and wouldn't require as much from the individual relative to going to in-person treatment, which is more ecologically valid and scalable.

8 CONCLUSION

The treatment needs and clinical status of individuals with AVH can vary and change over time [58]. Our study highlights the usefulness and feasibility of using voice diary EMA to assess the severity of AVH. Table 3 demonstrates that the complete model, which integrates audio, text from voice diaries, and mobile sensor data, achieved the highest level of performance. However, it is noteworthy that acceptable results can still be obtained by solely utilizing audio and manually transcribed text from voice diaries. Moreover, the automated process presents the possibility of developing an AVH predictive tool with minimal human intervention. The methodology presents new opportunities for the deployment of health-centric mobile sensing applications, offering a potentially more energy-efficient and privacy-conscious alternative. Furthermore, the successful online recruitment of participants in our study shows their willingness to use EMA and smartphone technology for self-tracking of AVH. Voice diary platforms possess the potential to assume a pivotal role in conducting comprehensive mental health evaluations in the future.

ACKNOWLEDGMENTS

This work is supported by National Institute of Mental Health (NIMH), grant number R01MH112641.

REFERENCES

[1] Daniel A Adler, Dror Ben-Zeev, Vincent WS Tseng, John M Kane, Rachel Brian, Andrew T Campbell, Marta Hauser, Emily A Scherer, and Tanzeem Choudhury. 2020. Predicting early warning signs of psychotic relapse from passive sensing data: an approach using encoder-decoder neural networks. *JMIR mHealth and uHealth* 8, 8 (2020), e19962.

[2] Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, et al. 2016. Deep speech 2: End-to-end speech recognition in english and mandarin. In *International conference on machine learning*. PMLR, 173–182.

[3] Nancy C Andreasen and Michael Flaum. 1991. Schizophrenia: the characteristic symptoms. *Schizophrenia bulletin* 17, 1 (1991), 27–49.

[4] Nancy C Andreasen and William M Grove. 1986. Thought, language, and communication in schizophrenia: diagnosis and prognosis. *Schizophrenia bulletin* 12, 3 (1986), 348–359.

[5] American Psychiatric Association. 2013. *Diagnostic and statistical manual of mental disorders: DSM-5*. Vol. 5. American psychiatric association Washington, DC.

[6] Min Hane Aung, Mark Matthews, and Tanzeem Choudhury. 2017. Sensing behavioral symptoms of mental health and delivering personalized interventions using mobile technologies. *Depression and anxiety* 34, 7 (2017), 603–609.

[7] Yi Ji Bae, Midan Shim, and Won Hee Lee. 2021. Schizophrenia detection using machine learning approach from social media content. *Sensors* 21, 17 (2021), 5924.

[8] Jakob E Bardram and Aleksandar Matic. 2020. A decade of ubiquitous computing research in mental health. *IEEE Pervasive Computing* 19, 1 (2020), 62–72.

[9] Robert H Belmaker and Galila Agam. 2008. Major depressive disorder. *New England Journal of Medicine* 358, 1 (2008), 55–68.

[10] Dror Ben-Zeev, Rachel Brian, Rui Wang, Weichen Wang, Andrew T Campbell, Min SH Aung, Michael Merrill, Vincent WS Tseng, Tanzeem Choudhury, Marta Hauser, et al. 2017. CrossCheck: Integrating self-report, behavioral sensing, and smartphone use to identify digital indicators of psychotic relapse. *Psychiatric rehabilitation journal* 40, 3 (2017), 266.

[11] Dror Ben-Zeev, Emily A Scherer, Rui Wang, Haiyi Xie, and Andrew T Campbell. 2015. Next-generation psychiatric assessment: Using smartphone sensors to monitor behavior and mental health. *Psychiatric rehabilitation journal* 38, 3 (2015), 218.

[12] Dror Ben-Zeev, Rui Wang, Saeed Abdullah, Rachel Brian, Emily A Scherer, Lisa A Mistler, Marta Hauser, John M Kane, Andrew Campbell, and Tanzeem Choudhury. 2016. Mobile behavioral sensing for outpatients and inpatients with schizophrenia. *Psychiatric services* 67, 5 (2016), 558–561.

[13] Josef Bless, Runar Smelror, Ingrid Agartz, and Kenneth Hugdahl. 2017. SA110. Using a Smartphone App to Assess Auditory Hallucinations in Adolescent Schizophrenia: Is This the Way to go for Better Control Over Voices? *Schizophrenia bulletin* 43, Suppl 1 (2017), S152.

[14] Mehdi Boukhechba, Yu Huang, Philip Chow, Karl Fua, Bethany A Teachman, and Laura E Barnes. 2017. Monitoring social anxiety from mobility and communication patterns. In *Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers*. 749–753.

[15] Vera Brink, Catheleine van Driel, Saliha El Bouhaddani, Klaas J Wardenaar, Lieke van Domburgh, Barbara Schaefer, Marije van Beilen, Agna A Bartels-Velthuis, and Wim Veling. 2020. Spontaneous discontinuation of distressing auditory verbal hallucinations in a school-based sample of adolescents: a longitudinal study. *European child & adolescent psychiatry* 29 (2020), 777–790.

[16] Xiao Chang, Yi-Bin Xi, Long-Biao Cui, Hua-Ning Wang, Jin-Bo Sun, Yuan-Qiang Zhu, Peng Huang, Guusje Collin, Kang Liu, Min Xi, et al. 2015. Distinct inter-hemispheric dysconnectivity in schizophrenia patients with and without auditory verbal hallucinations. *Scientific Reports* 5, 1 (2015), 1–12.

[17] Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. 785–794.

[18] Xingui Chen, Gong-Jun Ji, Chunyan Zhu, Xiaomeng Bai, Lu Wang, Kongliang He, Yaxiang Gao, Longxiang Tao, Fengqiong Yu, Yanghua Tian, et al. 2019. Neural correlates of auditory verbal hallucinations in schizophrenia and the therapeutic response to theta-burst transcranial magnetic stimulation. *Schizophrenia bulletin* 45, 2 (2019), 474–483.

[19] Zhenyu Chen, Mu Lin, Fanglin Chen, Nicholas D Lane, Giuseppe Cardone, Rui Wang, Tianxing Li, Yiqiang Chen, Tanzeem Choudhury, and Andrew T Campbell. 2013. Unobtrusive sleep monitoring using smartphones. In *Proceedings of the 7th International Conference on Pervasive Computing Technologies for Healthcare*. ICST (Institute for Computer Sciences, Social-Informatics and …, 145–152.

[20] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078* (2014).

[21] Michael A Cohn, Matthias R Mehl, and James W Pennebaker. 2004. Linguistic markers of psychological change surrounding September 11, 2001. *Psychological science* 15, 10 (2004), 687–693.

[22] Cheryl M Corcoran, Facundo Carrillo, Diego Fernández-Slezak, Gillinder Bedi, Casimir Klim, Daniel C Javitt, Carrie E Bearden, and Guillermo A Cecchi. 2018. Prediction of psychosis across protocols and risk cohorts using automated language analysis. *World Psychiatry* 17, 1 (2018), 67–75.

[23] H Corona-Hernández, SG Brederoo, JN de Boer, and IEC Sommer. 2022. A data-driven linguistic characterization of hallucinated voices in clinical and non-clinical voice-hearers. *Schizophrenia Research* 241 (2022), 210–217.
[24] Benjamin Sage Crosier, Rachel Marie Brian, and Dror Ben-Zeev. 2016. Using Facebook to reach people who experience auditory hallucinations. *Journal of medical Internet research* 18, 6 (2016), e160.
[25] Nicholas Cummins, Alice Baird, and Bjoern W Schuller. 2018. Speech analysis for health: Current state-of-the-art and the increasing impact of deep learning. *Methods* 151 (2018), 41–54.
[26] Bruce N Cuthbert et al. 2014. The RDoC framework: continuing commentary. *World Psychiatry* 13, 2 (2014), 196.
[27] Kirstin Daalman, Marco PM Boks, Kelly MJ Diederen, Antoin D de Weijer, Jan Dirk Blom, René S Kahn, and Iris EC Sommer. 2011. The same or different? A phenomenological comparison of auditory verbal hallucinations in healthy and psychotic individuals. *The Journal of clinical psychiatry* 72, 3 (2011), 0–0.
[28] K Daalman, IEC Sommer, EM Derks, and ER Peters. 2013. Cognitive biases and auditory verbal hallucinations in healthy and clinical individuals. *Psychological Medicine* 43, 11 (2013), 2339–2347.
[29] Saskia de Leede-Smith and Emma Barkus. 2013. A comprehensive review of auditory verbal hallucinations: lifetime prevalence, correlates and mechanisms in healthy and clinical individuals. *Frontiers in human neuroscience* 7 (2013), 367.
[30] Philippe Delespaul, Marten devries, and Jim van Os. 2002. Determinants of occurrence and recovery from hallucinations in daily life. *Social psychiatry and psychiatric epidemiology* 37, 3 (2002), 97–104.
[31] Sasha Deutsch-Link. 2016. Language in schizophrenia: What we can learn from quantitative text analysis. 2047 (2016).
[32] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).
[33] Clement Donde, David Luck, Stephanie Grot, David I Leitman, Jerome Brunelin, and Frederic Haesebaert. 2017. Tone-matching ability in patients with schizophrenia: A systematic review and meta-analysis. *Schizophrenia Research* 181 (2017), 94–99.
[34] Roisin Doyle, Niall Turner, Felicity Fanning, Daria Brennan, Laoise Renwick, Elizabeth Lawlor, and Mary Clarke. 2014. First-episode psychosis and disengagement from treatment: a systematic review. *Psychiatric Services* 65, 5 (2014), 603–611.
[35] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise.. In *Kdd*, Vol. 96. 226–231.
[36] Florian Eyben, Klaus R Scherer, Björn W Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y Devillers, Julien Epps, Petri Laukka, Shrikanth S Narayanan, et al. 2015. The Geneva minimalistic acoustic parameter set (GeMAPS) for voice research and affective computing. *IEEE transactions on affective computing* 7, 2 (2015), 190–202.
[37] Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*. 1459–1462.
[38] Florian Eyben, Martin Wöllmer, and Björn Schuller. 2020. openSMILE. https://github.com/audeering/opensmile.
[39] Denzil Ferreira, Vassilis Kostakos, and Anind K Dey. 2015. AWARE: mobile context instrumentation framework. *Frontiers in ICT* 2 (2015), 6.
[40] Judith M Ford. 2016. Studying auditory verbal hallucinations using the RDoC framework. *Psychophysiology* 53, 3 (2016), 298–304.
[41] William I Fraser, Kathleen M King, Philip Thomas, and Robert E Kendell. 1986. The diagnosis of schizophrenia by language analysis. *The British Journal of Psychiatry* 148, 3 (1986), 275–278.
[42] Daniel Freeman and Philippa A Garety. 2003. Connecting neurosis and psychosis: the direct influence of emotion on delusions and hallucinations. *Behaviour research and therapy* 41, 8 (2003), 923–947.
[43] Christopher D Frith and D John Done. 1988. Towards a neuropsychology of schizophrenia. *The British Journal of Psychiatry* 153, 4 (1988), 437–443.
[44] Kelvin MT Fung, Hector WH Tsang, and Patrick W Corrigan. 2008. Self-stigma of people with schizophrenia as predictor of their adherence to psychosocial treatment. *Psychiatric rehabilitation journal* 32, 2 (2008), 95.
[45] Google Activity Recognition Api. 2019. Google Activity Recognition Api. https://developers.google.com/android/reference/com/google/android/gms/location/ActivityRecognitionClient.
[46] Petra C Gronholm, Graham Thornicroft, Kristin R Laurens, and Sara Evans-Lacko. 2017. Mental health-related stigma and pathways to care for people at risk of psychotic disorders or experiencing first-episode psychosis: a systematic review. *Psychological medicine* 47, 11 (2017), 1867–1879.
[47] Agnes Grünerbl, Amir Muaremi, Venet Osmani, Gernot Bahle, Stefan Oehler, Gerhard Tröster, Oscar Mayora, Christian Haring, and Paul Lukowicz. 2014. Smartphone-based recognition of states and state changes in bipolar disorder patients. *IEEE journal of biomedical and health informatics* 19, 1 (2014), 140–148.
[48] Gillian Haddock, J McCarron, N Tarrier, and EB Faragher. 1999. Scales to measure dimensions of hallucinations and delusions: the psychotic symptom rating scales (PSYRATS). *Psychological medicine* 29, 4 (1999), 879–889.
[49] S Hartley, G Haddock, D Vasconcelos e Sa, R Emsley, and C Barrowclough. 2014. An experience sampling study of worry and rumination in psychosis. *Psychological Medicine* 44, 8 (2014), 1605–1614.

[50] Nik Wahidah Hashim, Mitch Wilkes, Ronald Salomon, Jared Meggs, and Daniel J France. 2017. Evaluation of voice acoustics as predictors of clinical depression scores. *Journal of Voice* 31, 2 (2017), 256–e1.

[51] Karl Herholz, Alexander Thiel, Klaus Wienhard, Uwe Pietrzyk, H-M Von Stockhausen, Hans Karbe, J Kessler, Thomas Bruckbauer, Marco Halber, and W-D Heiss. 1996. Individual functional anatomy of verb generation. *Neuroimage* 3, 3 (1996), 185–194.

[52] RE Hoffman, M Varanko, J Gilmore, and AL Mishara. 2008. Experiential features used by patients with schizophrenia to differentiate 'voices' from ordinary verbal thought. *Psychological medicine* 38, 8 (2008), 1167–1176.

[53] Ralph E Hoffman. 1986. Verbal hallucinations and language production processes in schizophrenia. *Behavioral and Brain Sciences* 9, 3 (1986), 503–517.

[54] Ralph E Hoffman. 2007. A social deafferentation hypothesis for induction of active schizophrenia. *Schizophrenia bulletin* 33, 5 (2007), 1066–1070.

[55] Daniel C Javitt. 2009. When doors of perception close: bottom-up models of disrupted cognition in schizophrenia. *Annual review of clinical psychology* 5 (2009), 249–275.

[56] Daniel C Javitt and Robert A Sweet. 2015. Auditory dysfunction in schizophrenia: integrating clinical and basic features. *Nature Reviews Neuroscience* 16, 9 (2015), 535–550.

[57] Louise C. Johns, Mary Cannon, Nicola Singleton, Robin M. Murray, Michael Farrell, Traolach Brugha, Paul Bebbington, Rachel Jenkins, and Howard Meltzer. 2004. Prevalence and correlates of self-reported psychotic symptoms in the British population. *British Journal of Psychiatry* 185, 4 (Oct. 2004), 298–305. https://doi.org/10.1192/bjp.185.4.298

[58] Louise C Johns, Kristiina Kompus, Melissa Connell, Clara Humpston, Tania M Lincoln, Eleanor Longden, Antonio Preti, Ben Alderson-Day, Johanna C Badcock, Matteo Cella, et al. 2014. Auditory verbal hallucinations in persons with and without a need for care. *Schizophrenia bulletin* 40, Suppl_4 (2014), S255–S264.

[59] Louise C Johns, James Y Nazroo, Paul Bebbington, and Elizabeth Kuipers. 2002. Occurrence of hallucinatory experiences in a community sample and ethnic variations. *The British Journal of Psychiatry* 180, 2 (2002), 174–178.

[60] Ewa Kacewicz, James W Pennebaker, Matthew Davis, Moongee Jeon, and Arthur C Graesser. 2013. Pronoun use reflects standings in social hierarchies. *Journal of Language and Social Psychology* (2013), 0261927X13502654.

[61] Se Hyun Kim, Hee Yeon Jung, Samuel S Hwang, Jae Seung Chang, Yeni Kim, Yong Min Ahn, and Yong Sik Kim. 2010. The usefulness of a self-report questionnaire measuring auditory verbal hallucinations. *Progress in Neuro-Psychopharmacology and Biological Psychiatry* 34, 6 (2010), 968–973.

[62] David Kimhy, Melanie M Wall, Marie C Hansen, Julia Vakhrusheva, C Jean Choi, Philippe Delespaul, Nicholas Tarrier, Richard P Sloan, and Dolores Malaspina. 2017. Autonomic Regulation and Auditory Hallucinations in Individuals With Schizophrenia: An Experience Sampling Study. *Schizophrenia Bulletin* 43, 4 (Feb. 2017), 754–763. https://doi.org/10.1093/schbul/sbw219

[63] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[64] Mrinal Kumar, Mark Dredze, Glen Coppersmith, and Munmun De Choudhury. 2015. Detecting changes in suicide content manifested in social media following celebrity suicides. In *Proceedings of the 26th ACM Conference on Hypertext & Social Media*. ACM, 85–94.

[65] Gina R Kuperberg, Philip K McGuire, Edward T Bullmore, Michael J Brammer, Sophie Rabe-Hesketh, Ian C Wright, David J Lythgoe, Steven CR Williams, and Anthony S David. 2000. Common and distinct neural substrates for pragmatic, semantic, and syntactic processing of spoken sentences: an fMRI study. *Journal of Cognitive Neuroscience* 12, 2 (2000), 321–341.

[66] Frank Larøi, Iris E Sommer, Jan Dirk Blom, Charles Fernyhough, Dominic H Ffytche, Kenneth Hugdahl, Louise C Johns, Simon McCarthy-Jones, Antonio Preti, Andrea Raballo, et al. 2012. The characteristic features of auditory verbal hallucinations in clinical and nonclinical groups: state-of-the-art overview and future directions. *Schizophrenia bulletin* 38, 4 (2012), 724–733.

[67] Josephine Lau, Benjamin Zimmerman, and Florian Schaub. 2018. Alexa, are you listening? Privacy perceptions, concerns and privacy-seeking behaviors with smart speakers. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–31.

[68] Belinda R Lennox, S Bert, G Park, Peter B Jones, and Peter G Morris. 1999. Spatial and temporal mapping of neural activity associated with auditory hallucinations. *The Lancet* 353, 9153 (1999), 644.

[69] Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. 2017. A structured self-attentive sentence embedding. *arXiv preprint arXiv:1703.03130* (2017).

[70] Tania M Lincoln, Winfried Rief, Stefan Westermann, Michael Ziegler, Marie-Luise Kesting, Eva Heibach, and Stephanie Mehl. 2014. Who stays, who benefits? Predicting dropout and change in cognitive behaviour therapy for psychosis. *Psychiatry Research* 216, 2 (2014), 198–205.

[71] Hong Lu, Denise Frauendorfer, Mashfiqui Rabbi, Marianne Schmid Mast, Gokul T Chittaranjan, Andrew T Campbell, Daniel Gatica-Perez, and Tanzeem Choudhury. 2012. Stresssense: Detecting stress in unconstrained acoustic environments using smartphones. In *Proceedings of the 2012 ACM conference on ubiquitous computing*. 351–360.

[72] Scott M Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.). Curran Associates, Inc., 4765–4774. http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf

[73] Masking and padding with Keras. 2021. Masking and padding with Keras. https://www.tensorflow.org/guide/keras/masking_and_padding.
[74] John McGrath, Sukanta Saha, Joy Welham, Ossama El Saadi, Clare MacCauley, and David Chant. 2004. A systematic review of the incidence of schizophrenia: the distribution of rates and the influence of sex, urbanicity, migrant status and methodology. *BMC medicine* 2, 1 (2004), 1–22.
[75] Colette M McKay, Donna M Headlam, and David L Copolov. 2000. Central auditory processing in patients with auditory hallucinations. *American Journal of Psychiatry* 157, 5 (2000), 759–766.
[76] Neil M McLachlan, Dougal S Phillips, Susan L Rossell, and Sarah J Wilson. 2013. Auditory processing and hallucinations in schizophrenia. *Schizophrenia research* 150, 2-3 (2013), 380–385.
[77] Emiliano Miluzzo, Nicholas D Lane, Shane B Eisenman, and Andrew T Campbell. 2007. CenceMe–injecting sensing presence into social networking applications. In *European Conference on Smart Sensing and Context*. Springer, 1–28.
[78] Kyle S Minor, Beshaun J Davis, Matthew P Marggraf, Lauren Luther, and Megan L Robbins. 2018. Words matter: Implementing the electronically activated recorder in schizotypy. *Personality Disorders: Theory, Research, and Treatment* 9, 2 (2018), 133.
[79] Shayan Mirjafari, Kizito Masaba, Ted Grover, Weichen Wang, Pino Audia, Andrew T Campbell, Nitesh V Chawla, Vedant Das Swain, Munmun De Choudhury, Anind K Dey, et al. 2019. Differentiating higher and lower job performers in the workplace using mobile sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 3, 2 (2019), 1–24.
[80] David C Mohr, Mi Zhang, and Stephen M Schueller. 2017. Personal sensing: understanding mental health using ubiquitous sensors and machine learning. *Annual review of clinical psychology* 13 (2017), 23–47.
[81] Rodney Morice and Don McNicol. 1986. Language changes in schizophrenia: a limited replication. *Schizophrenia Bulletin* 12, 2 (1986), 239–251.
[82] Isaac Moshe, Yannik Terhorst, Kennedy Opoku Asare, Lasse Bosse Sander, Denzil Ferreira, Harald Baumeister, David C Mohr, and Laura Pulkki-Råback. 2021. Predicting Symptoms of Depression and Anxiety Using Smartphone and Wearable Data. *Frontiers in psychiatry* 12 (2021).
[83] Amir Muaremi, Franz Gravenhorst, Agnes Grünerbl, Bert Arnrich, and Gerhard Tröster. 2014. Assessing bipolar episodes using speech cues derived from phone calls. In *Pervasive Computing Paradigms for Mental Health: 4th International Symposium, MindCare 2014, Tokyo, Japan, May 8-9, 2014, Revised Selected Papers 4*. Springer, 103–114.
[84] Matthew L Newman, James W Pennebaker, Diane S Berry, and Jane M Richards. 2003. Lying words: Predicting deception from linguistic styles. *Personality and social psychology bulletin* 29, 5 (2003), 665–675.
[85] Stefanie Nickels, Matthew D Edwards, Sarah F Poole, Dale Winter, Jessica Gronsbell, Bella Rozenkrants, David P Miller, Mathias Fleck, Alan McLean, Bret Peterson, et al. 2021. Toward a Mobile Platform for Real-world Digital Measurement of Depression: User-Centered Design, Data Quality, and Behavioral and Clinical Modeling. *JMIR mental health* 8, 8 (2021), e27589.
[86] Jukka-Pekka Onnela, Caleb Dixon, Keary Griffin, Tucker Jaenicke, Leila Minowada, Sean Esterkin, Alvin Siu, Josh Zagorsky, and Eli Jones. 2021. Beiwe: A data collection platform for high-throughput digital phenotyping. *Journal of Open Source Software* 6, 68 (2021), 3417.
[87] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in PyTorch. (2017).
[88] Paola Pedrelli, Szymon Fedor, Asma Ghandeharioun, Esther Howe, Dawn F Ionescu, Darian Bhathena, Lauren B Fisher, Cristina Cusin, Maren Nyer, Albert Yeung, et al. 2020. Monitoring changes in depression severity using wearable and mobile sensors. *Frontiers in psychiatry* 11 (2020), 1413.
[89] James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. The development and psychometric properties of LIWC2015. *UT Faculty/Researcher Works* (2015).
[90] James W Pennebaker, Cindy K Chung, Joey Frazee, Gary M Lavergne, and David I Beaver. 2014. When small words foretell academic success: The case of college admissions essays. *PloS one* 9, 12 (2014), e115844.
[91] James W Pennebaker, Matthias R Mehl, and Kate G Niederhoffer. 2003. Psychological aspects of natural language use: Our words, our selves. *Annual review of psychology* 54, 1 (2003), 547–577.
[92] Viliam Rapcan, Shona D'Arcy, Sherlyn Yeap, Natasha Afzal, Jogin Thakore, and Richard B Reilly. 2010. Acoustic and temporal analysis of speech: A potential biomarker for schizophrenia. *Medical engineering & physics* 32, 9 (2010), 1074–1079.
[93] Benjamin Rolland, Ali Amad, Emmanuel Poulet, Régis Bordet, Alexandre Vignaud, Rémy Bation, Christine Delmaire, Pierre Thomas, Olivier Cottencin, and Renaud Jardri. 2015. Resting-state functional connectivity of the nucleus accumbens in auditory and visual hallucinations in schizophrenia. *Schizophrenia bulletin* 41, 1 (2015), 291–299.
[94] Matthia Sabatelli, Venet Osmani, Oscar Mayora, Agnes Gruenerbl, and Paul Lukowicz. 2014. Correlation of significant places with self-reported state of bipolar disorder patients. In *2014 4th International Conference on Wireless Mobile Communication and Healthcare-Transforming Healthcare Through Innovations in Mobile and Wireless Technologies (MOBIHEALTH)*. IEEE, 116–119.
[95] Norihiro Sadato, Yoshiharu Yonekura, Hiroki Yamada, Satoshi Nakamura, Atsuo Waki, and Yasushi Ishii. 1998. Activation patterns of covert word generation detected by fMRI: comparison with 3D PET. *Journal of computer assisted tomography* 22, 6 (1998), 945–952.

[96] Sohrab Saeb, Emily G Lattie, Konrad P Kording, and David C Mohr. 2017. Mobile phone detection of semantic location and its relationship to depression and anxiety. *JMIR mHealth and uHealth* 5, 8 (2017), e7297.

[97] Sohrab Saeb, Mi Zhang, Christopher J Karr, Stephen M Schueller, Marya E Corden, Konrad P Kording, and David C Mohr. 2015. Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: an exploratory study. *Journal of medical Internet research* 17, 7 (2015).

[98] Koustuv Saha, Ted Grover, Stephen M Mattingly, Vedant Das Swain, Pranshu Gupta, Gonzalo J Martinez, Pablo Robles-Granda, Gloria Mark, Aaron Striegel, and Munmun De Choudhury. 2021. Person-Centered Predictions of Psychological Constructs with Social Media Contextualized by Multimodal Sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 5, 1 (2021), 1–32.

[99] Shekhar Saxena, Graham Thornicroft, Martin Knapp, and Harvey Whiteford. 2007. Resources for mental health: scarcity, inequity, and inefficiency. *The lancet* 370, 9590 (2007), 878–889.

[100] Terrence J Sejnowski. 2018. *The deep learning revolution*. MIT press.

[101] Bryony Sheaves, Paul E Bebbington, Guy M Goodwin, Paul J Harrison, Colin A Espie, Russell G Foster, and Daniel Freeman. 2016. Insomnia and hallucinations in the general population: findings from the 2000 and 2007 British Psychiatric Morbidity Surveys. *Psychiatry Research* 241 (2016), 141–146.

[102] Jessica Helen Silver, Marcus Lewton, and Heledd Wyn Lewis. 2023. Mediators of negative content and voice-related distress in a diverse sample of clinical and non-clinical voice-hearers. *British Journal of Clinical Psychology* 62, 1 (2023), 96–111.

[103] Robert R Sinclair and Janelle H Cheung. 2016. Money matters: Recommendations for financial stress research in occupational health psychology. *Stress and Health* 32, 3 (2016), 181–193.

[104] Runar Elle Smelror, Josef Johann Bless, Kenneth Hugdahl, and Ingrid Agartz. 2019. Feasibility and Acceptability of Using a Mobile Phone App for Characterizing Auditory Verbal Hallucinations in Adolescents With Early-Onset Psychosis: Exploratory Study. *JMIR Formative Research* 3, 2 (May 2019), e13882. https://doi.org/10.2196/13882

[105] Iris EC Sommer, Kirstin Daalman, Thomas Rietkerk, Kelly M Diederen, Steven Bakker, Jaap Wijkstra, and Marco PM Boks. 2010. Healthy individuals with auditory verbal hallucinations; who are they? Psychiatric assessments of a selected sample of 103 subjects. *Schizophrenia bulletin* 36, 3 (2010), 633–641.

[106] Iris EC Sommer, Kelly MJ Diederen, Jan-Dirk Blom, Anne Willems, Leila Kushan, Karin Slotema, Marco PM Boks, Kirstin Daalman, Hans W Hoek, Sebastiaan FW Neggers, et al. 2008. Auditory verbal hallucinations predominantly activate the right inferior frontal area. *Brain* 131, 12 (2008), 3169–3177.

[107] M Stephane, S Barton, and NN Boutros. 2001. Auditory verbal hallucinations and dysfunction of the neural substrates of speech. *Schizophrenia research* 50, 1-2 (2001), 61–78.

[108] Rael D Strous, Nelson Cowan, Walter Ritter, and Daniel C Javitt. 1995. Auditory sensory (" echoic") memory dysfunction in schizophrenia. *The American journal of psychiatry* (1995).

[109] Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of language and social psychology* 29, 1 (2010), 24–54.

[110] A. Y. Tien. 1991. Distribution of hallucinations in the population. *Social Psychiatry and Psychiatric Epidemiology* 26, 6 (1991), 287–292. https://doi.org/10.1007/bf00789221

[111] Vincent W-S Tseng, Akane Sano, Dror Ben-Zeev, Rachel Brian, Andrew T Campbell, Marta Hauser, John M Kane, Emily A Scherer, Rui Wang, Weichen Wang, et al. 2020. Using behavioral rhythms and multi-task learning to predict fine-grained symptoms of schizophrenia. *Scientific reports* 10, 1 (2020), 1–17.

[112] Rachel Tucker, John Farhall, Neil Thomas, Christopher Groot, and Susan L Rossell. 2013. An examination of auditory processing and affective prosody in relatives of patients with auditory hallucinations. *Frontiers in Human Neuroscience* 7 (2013), 531.

[113] Ryan J Van Lieshout and Joel O Goldberg. 2007. Quantifying self-reports of auditory verbal hallucinations in persons with psychosis. *Canadian Journal of Behavioural Science/Revue canadienne des sciences du comportement* 39, 1 (2007), 73.

[114] Fabian Wahle, Tobias Kowatsch, Elgar Fleisch, Michael Rufer, Steffi Weidt, et al. 2016. Mobile sensing and support for people with depression: a pilot trial in the wild. *JMIR mHealth and uHealth* 4, 3 (2016), e5960.

[115] Rui Wang, Fanglin Chen, Zhenyu Chen, Tianxing Li, Gabriella Harari, Stefanie Tignor, Xia Zhou, Dror Ben-Zeev, and Andrew T Campbell. 2014. StudentLife: assessing mental health, academic performance and behavioral trends of college students using smartphones. In *Proceedings of the 2014 ACM international joint conference on pervasive and ubiquitous computing*. 3–14.

[116] Rui Wang, Min S. H. Aung, Saeed Abdullah, Rachel Brian, Andrew T. Campbell, Tanzeem Choudhury, Martan Hauser, John Kane, Michael Merrill, Emily A. Scherer, and Vincent W. S. Tseng. 2016. CrossCheck: Toward passive sensing and detection of mental health changes in people with schizophrenia. (2016).

[117] Rui Wang, Weichen Wang, Min SH Aung, Dror Ben-Zeev, Rachel Brian, Andrew T Campbell, Tanzeem Choudhury, Marta Hauser, John Kane, Emily A Scherer, et al. 2017. Predicting symptom trajectories of schizophrenia using mobile sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 1, 3 (2017), 1–24.

[118] Rui Wang, Weichen Wang, Alex DaSilva, Jeremy F Huckins, William M Kelley, Todd F Heatherton, and Andrew T Campbell. 2018. Tracking depression dynamics in college students using mobile phone and wearable sensing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 1 (2018), 1–26.

[119] Weichen Wang, Gabriella M Harari, Rui Wang, Sandrine R Müller, Shayan Mirjafari, Kizito Masaba, and Andrew T Campbell. 2018. Sensing behavioral change over time: Using within-person variability features from mobile sensing to predict personality traits. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 3 (2018), 1–21.

[120] Weichen Wang, Shayan Mirjafari, Gabriella Harari, Dror Ben-Zeev, Rachel Brian, Tanzeem Choudhury, Marta Hauser, John Kane, Kizito Masaba, Subigya Nepal, et al. 2020. Social sensing: assessing social functioning of patients living with schizophrenia using mobile phone sensing. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–15.

[121] Weichen Wang, Subigya Nepal, Jeremy F. Huckins, Lessley Hernandez, Vlado Vojdanovski, Dante Mack, Jane Plomp, Arvind Pillai, Mikio Obuchi, Alex daSilva, Eilis Murphy, Elin Hedlund, Courtney Rogers, Meghan Meyer, and Andrew Campbell. 2022. First-Gen Lens: Assessing Mental Health of First-Generation Students across Their First Year at College Using Mobile Sensing. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 6, 2, Article 95 (jul 2022), 32 pages. https://doi.org/10.1145/3543194

[122] Flavie Waters, Daniel Collerton, Dominic H Ffytche, Renaud Jardri, Delphine Pins, Robert Dudley, Jan Dirk Blom, Urs Peter Mosimann, Frank Eperjesi, Stephen Ford, et al. 2014. Visual hallucinations in the psychosis spectrum and comparative information from neurodegenerative disorders and eye disease. *Schizophrenia bulletin* 40, Suppl_4 (2014), S233–S245.

[123] Danny Wyatt, Tanzeem Choudhury, and Jeff A Bilmes. 2007. Conversation detection and speaker segmentation in privacy-sensitive situated speech data.. In *INTERSPEECH*. 586–589.

[124] Danny Wyatt, Tanzeem Choudhury, Jeff A Bilmes, and Henry A Kautz. 2007. A Privacy-Sensitive Approach to Modeling Multi-Person Conversations.. In *IJCAI*, Vol. 7. 1769–1775.

[125] Danny Wyatt, Tanzeem Choudhury, and Henry Kautz. 2007. Capturing spontaneous conversation and social dynamics: A privacy-sensitive data collection effort. In *Acoustics, Speech and Signal Processing, 2007. ICASSP 2007. IEEE International Conference on*, Vol. 4. IEEE, IV–213.

[126] Weizhe Xu, Jake Portanova, Ayesha Chander, Dror Ben-Zeev, and Trevor Cohen. 2020. The Centroid Cannot Hold: Comparing Sequential and Global Estimates of Coherence as Indicators of Formal Thought Disorder. In *AMIA Annual Symposium Proceedings*, Vol. 2020. American Medical Informatics Association, 1315.

[127] Weizhe Xu, Weichen Wang, Jake Portanova, Ayesha Chander, Andrew Campbell, Serguei Pakhomov, Dror Ben-Zeev, and Trevor Cohen. 2022. Fully Automated Detection of Formal Thought Disorder with Time-series Augmented Representations for Detection of Incoherent Speech (TARDIS). *Journal of Biomedical Informatics* (2022), 103998.