Sociability impairment, such as decreased social network size and socialization, is implicated in mental health disorders. To complement the existing self-reports-based assessment of sociability measures, which could be error-prone and burdensome, we propose to estimate an individual’s everyday conversational network from free-living speech recordings obtained with a wearable. Our first contribution is ECoNet, an automatic method to estimate the everyday conversational network using a modular audio processing architecture. Our second contribution is using ECoNet to analyze multiday egocentric audio recordings from 32 individuals representing diverse mental health conditions (healthy controls, depressive disorders, and psychotic disorders). Specifically, we discover that the conversational network size as a sociability measure has a significant correlation with mental health scores. For example, the correlation coefficient between network size and depression severity score was $-0.56 \ (p < 0.01)$. Audio-based estimation of conversational network size using ECoNet, therefore, could provide a pervasive computing solution to complement existing mental health assessment methods.
The speaker-counting task has primarily to assess ‘knowledge’ however, RFID tag-based methods are not practically viable in most cases, as tags are required on all individuals in one’s network. Thus, in this article, we focus on speech-based sensing.

Everyday conversational network: We define the everyday conversational network as the network of conversation partners appearing in egocentric audio recordings from a day. The audio recordings are expected to be multiple days long, thereby providing a longitudinal view of the individual’s network. Such recordings are now conveniently possible using inexpensive wristband audio recorders of smartwatch form factors. If we could identify and assign each utterance in recorded speech with a speaker label, we could estimate various network properties; the assigned speaker labels need not be the real identity of a person but could just be internal labels to distinguish unique speakers. The network’s size, referred to as the conversational network size in this work, is one such property that could be potentially useful as a sociability measure in mental health applications. Network size is commonly assessed, subjectively via self-report, for its potential relation with mental health.1,3–5

Related work: Conversational network size estimation is closely related to the audio-based speaker-counting task.11,12 The speaker-counting task has primarily focused on estimating the number of conversation partners in a short audio recording. However, conversational network size estimation requires inferring the number of conversation partners from multiday noise-filled recordings, with potentially repeated encounters with the same speakers. An audio processing pipeline that addresses the possible audio environment changes during the day and higher computational costs in processing long audio recordings is also required. Xu et al.11 proposed Crowd++, a speaker-counting system using mel-frequency cepstral coefficient vector representation of speech segments. Variation of Crowd++ using different speech representations has also been proposed.12 The Crowd++ system was used in Xu et al.’s work13 to assess an individual’s daily interactions. However, multiple encounters with the same speaker were not tracked and speaker-count’s relation with mental health was not investigated. A related task to speaker-counting is concurrent speaker count estimation, i.e., estimating the maximum number of overlapped speakers in a short audio segment (e.g., Stöter et al.’s work14). Concurrent speaker count models, however, have no individual speaker models and, thus, cannot track speakers over time for conversational network size estimation.

Key challenges: The key challenge in estimating everyday conversational networks in a privacy-preserving manner is the absence of a priori knowledge about potentially encountered speakers, specifically, the absence of their speech samples or knowledge about when and how often they will be encountered. Even in a noiseless environment, identifying unique speakers can be challenging, e.g., when two speakers have similar voice characteristics. In our case, the challenges are compounded by free-living recording originating in arbitrarily noisy environments.

Main contributions: Our contributions in this article are twofold. First, we develop an automatic and robust estimation pipeline, labeled ECoNet, that estimates the unique speakers and their timing of speech from a day’s audio recording. Our key innovation is an unsupervised speaker identification using speaker-discriminative speech segment representation along with a machine learning model for spurious speaker detection. The key insight in developing a spurious speaker detection model is that the duration and occurrence of the spurious speakers’ speech segments might be different (more random) than those of true speakers. ECoNet has a 14.5% lower speaker-count error compared with the existing speaker-counting system. ECoNet has a modular architecture and uses block-based audio processing with speaker tracking to reduce computation and manage environment heterogeneity. We leveraged diverse public datasets for training ECoNet, and no speech content is parsed for privacy preservation. We note an important limitation of ECoNet, in that the assessed network does not distinguish who is talking to whom, and thus, whether the target individual (using the wearable) is involved in the conversation. We discuss the implications of this limitation in the “Interpretation” section.

Our second contribution is evaluating the potential use of conversational network size in mental health. We evaluated ECoNet in a pilot study acquiring multi-day egocentric audio recordings from 32 participants comprising healthy controls and participants with depressive and psychotic disorders. Our results show a significant association between conversational network size, estimated using ECoNet, and mental health scores, such as depression severity. Thus, our findings
imply that the conversational network size could serve as a complementary objective sociability measure, potentially useful for monitoring network size and engagement fluctuations in light of mental symptom changes.

ECONET: EVERYDAY CONVERSATIONAL NETWORK ESTIMATOR

In this section, we describe our proposed everyday conversational network estimator pipeline ECoNet. We adopted a modular architecture instead of a monolithic end-to-end architecture. The modular approach allowed us to leverage existing models and available well-labeled datasets to match our application. The architectural steps are demonstrated in Figure 1. To address the key challenges in everyday conversational network estimation, Steps 3–5 in Figure 1 depict our proposed approach to robustly detecting unique speakers in challenging audio environments.

ECoNet builds on the premise that an individual’s speech can provide a unique signature of the individual. In the absence of labeled speech samples from speakers, an unsupervised clustering with speaker-discriminative speech segment representation could potentially be used to identify the speakers. Any spuriously detected speakers due to noise and speech variabilities (e.g., speaking rate, style, etc.), however, should be filtered out to correctly assess key network parameters, such as the conversational network size. Higher computational costs and environmental heterogeneity in processing day-long audio recordings should also be managed, for which we propose block-based audio processing with speaker tracking.

ECoNet has five key steps in its pipeline, i.e., speech segment detection, speaker-embedding generation, clustering, filtering spuriously detected speakers, and speaker tracking.

**FIGURE 1.** ECoNet, the proposed everyday conversational network estimator pipeline. Speech segments in an audio block are detected with VAD. Speaker embeddings (X-vectors) are then extracted from the speech segments followed by embedding scoring and clustering, where each identified cluster represents a locally detected speaker. Spuriously detected speakers are removed and speakers across audio blocks are tracked to obtain a global set of speakers.

**Step 1: Speech Segment Detection**

The first step in estimating the everyday conversational network is to detect all the speech segments with a voice activity detection (VAD) model. We evaluated multiple VAD models, including recently proposed deep neural network (DNN)-based VAD models, to identify the best model for ECoNet. Liaqat et al.\(^{15}\) found existing VAD models to be highly erroneous for processing free-living audio, but DNN-based VAD models were not evaluated in their work. We employed the commonly used equal error rate (EER), the error rate where false-negative rate equals false positive rate, to compare VAD models. The PyanNet-architecture-based DNN model\(^{16,17}\) had a significantly low EER and was selected for ECoNet. The EER were 0.05 and 0.12 in the DIHARD evaluation dataset and an in-house validation dataset, respectively, representing a 77.2% and 43.3% improvement over the next best model (refer to Table S1 in the supplementary material, which is available in the IEEE Computer Society Digital Library at http://doi.ieeecomputer.org/10.1109/MPRV.2022.3155698).

**Step 2: Speaker Embeddings**

Speech segments could be represented by speaker embedding and clustered to assign a unique speaker label to segments from the same speaker. Speaker embedding is a speaker-discriminative speech representation, obtained as activations from a speaker...
identification DNN model’s final layers. Speech segments from the same speaker have a smaller interembedding distance than those from different speakers. We used time-delay neural network-based X-vectors\(^1\) as the speaker embedding. X-vectors have given state-of-the-art results across speaker identification and verification tasks\(^1\) but have not been explored for conversational network size estimation. X-vectors were computed using the recipe of Ryant et al.’s work\(^7\) (derived from Snyder et al.’s work\(^1\)).

**Step 3: Embedding Scoring and Clustering**

We computed the similarity between X-vectors (embedding scoring) using a probabilistic linear discriminant analysis (PLDA) model (see Section S-I-B in the supplementary material for details, available online). The similarity scores were clustered with agglomerative hierarchical clustering (AHC). The PLDA-based scoring for AHC clustering is commonly used for speaker recognition.\(^1\) With clustering, we assign a speaker label—corresponding to each identified cluster—to each of the speech segments.

**Step 4: Spurious Speaker Detection**

Noise and speech variability in free-living audio can cause spurious speaker detection, compromising the assessment of key everyday conversational network parameters. Spurious speakers are false positives in speaker detection, a new speaker assigned to a non-speech, or an existing speaker’s speech segment. A new speaker could also be assigned for overlapped speeches with significantly different embeddings from the individual speaker’s embeddings. Xu et al.\(^1\) used an interspeaker distance threshold to identify spurious speakers. However, a single threshold might not represent all spurious speaker detection scenarios. Toward a more representative model, we trained a random forest model using various speech-segment-based features. We hypothesized that the spuriously detected speakers would have different (likely random) features compared with the true speakers in a conversation. The trained model had an area under the ROC curve (AUROC) of 0.96 in the DIHARD evaluation dataset, indicating good detection performance (see Figure S1 in the supplementary material, available online). With the spurious speaker-detection model, ECoNet would filter falsely detected speakers before further evaluations (details of the model and speaker filtering are presented in Section S-I-C in the supplementary material, available online).

**Step 5: Speaker Tracking**

Steps 1–4 of ECoNet are applied in a smaller audio block to reduce computational costs and manage environment heterogeneity. To identify the same speaker across audio blocks of a given day, we propose clustering-based speaker tracking. We obtained speaker representation of each local speaker of a block, based on their embeddings, and clustered local speakers across the audio blocks to obtain a global speaker label (in the scope of a day) for the local speakers. We employed averaged-embedding-based speaker representation that gave the best clustering-based speaker tracking results in the DIHARD dataset (see Table S3 and Section S-I-F in the supplementary material for details, available online).

The global speakers identified in a day’s audio recording and their associated conversation properties constitute the inferred everyday conversational network. Its size, the conversational network size, represents the number of identified global speakers. Conversational network size is normalized by the audio duration and scaled by a reference duration to obtain a normalized metric comparable across days and individuals.

**Evaluation of ECoNet for Speaker-Counting Task**

We evaluated the performance of ECoNet to estimate the conversational network size, which we estimate by identifying unique speakers present in the recording; this metric is used in the “Sociability and Mental Health: A Pilot Study” section. We evaluated ECoNet for speaker-counting in the DIHARD evaluation set\(^17\) consisting of diverse audio recordings, such as from restaurants, meetings, clinical, and web videos. We used average error count distance (AECD), \[ AECD = \frac{1}{N} \sum_{i=1}^{N} (C_i - \hat{C}_i) \] with \(C_i\) and \(\hat{C}_i\) the true and estimated speaker count, and correlation coefficient between \(C_i\) and \(\hat{C}_i\) as the evaluation metric. ECoNet outperformed Crowd++,\(^11\) as well as its recently proposed extension,\(^12\) for speaker counting. AECD reduced by 14.15% and correlation improved from 0.24 to 0.70 using ECoNet compared with Crowd++ (refer to Table S2 in the supplementary material for evaluation results, available online). We also evaluated if ECoNet could detect group differences in the number of unique speakers present. Two groups of sequences from the DIHARD evaluation set were created with a significantly different number of speakers, testing five different grouping settings for evaluation. ECoNet was able to detect the difference when other speaker-counting systems were unable (see Figure S2 in the supplementary material, available online).
SOCIABILITY AND MENTAL HEALTH: A PILOT STUDY

An individual’s sociability measures could be computed from their everyday conversational network. One such sociability measure, related to the individual’s social network size implicated in mental illness, is the conversational network size. In this work, we evaluated if the conversational network size assessed using ECoNet is related to an individual’s mental health with a pilot study.

Study Design

Participants were recruited via a pilot study at the psychiatry clinic of Baylor College of Medicine (BCM)/Harris Health System, Houston, Texas, USA with 32 participants consisting of 11 outpatients with depression (depression group), eight outpatients with schizophrenia/schizoaffective disorders (psychosis group), and 13 healthy controls (healthy group). The healthy group was age-matched with the depression and psychosis group. We included a heterogeneous group in our pilot study to map sociability across disorders, aiming for sociability assessment as a transdiagnostic construct. The depression and psychosis groups were selected as these are the two most prevalent mental illnesses in the psychiatric clinics of BCM/Harris Health System. The study was triple approved by the IRB at Baylor College of Medicine, Rice University, and Harris Health System. Continuous audio recordings (suggested time of 8 AM to 8 PM) were obtained from participants in their free living, for up to seven days, using a wristband audio recorder (refer Sections S-II-A and S-II-B in the supplementary material for details, available online, on participant recruitment and summary of the obtained clinical audio dataset, respectively). The participants also reported the number of persons they interacted with, on each day of the study, in one of the ranges: 0–5, 6–10, 11–15, and >15. In our study, the wristband device added random date offsets on some recording days. Further, we had a differing number of days of self-reports and audio recording days for some subjects, e.g., if audio bands were not worn but self-reports provided. Thus, one-to-one mapping of all the audio recordings to their corresponding self-reports was not possible.

Mental Health and Personality Scores

Several questionnaire-based mental health and personality scores of the participants were obtained, and the relation of the conversational network size with these scores was analyzed. We obtained general measures, such as anxiety and depression scores, representing some of the prototypical fluctuations manifested when patients do better or worse clinically (e.g., feeling anxious before a social interaction). In future studies, we will include more symptom-specific measures also, such as those assessing psychotic features. We also obtained some trait-like scores (e.g., personality) to evaluate their relationship with participants’ conversational/social behavior. Such evaluation could inform future studies investigating multiple factors influencing one’s sociability. In particular, the following questionnaire-based scores were obtained.

- Depression: Patient health questionnaire (PHQ)-9 score.
- Anxiety: Generalized anxiety disorder (GAD)-7 questionnaire.
- Perceived stress scale (PSS) score.
- Distress tolerance scale (DTS) scores.
- Personality scores from the mini-IPIP questionnaire.

Statistical Test and Significance

We assessed group differences with a t-test (or Mann–Whitney test) for variables following (or not following) a normal distribution. We applied Bonferroni correction in multiple comparisons. The statistical significance was set as follows: * for $p < 0.05$, ** for $p < 0.01$, and *** for $p < 0.001$, where $p$ represents $p$-value.

Conversational Network Size and Mental Health

The ECoNet was used to assess the relationship between conversational network size and mental health in the clinical audio dataset. We set the audio block duration in the block-based processing of ECoNet to 2 hours and included days with at least 5 hours of audio for analysis.

Group Differences in the Estimated Conversational Network Size

We analyzed the conversational network size differences in three groups represented in our clinical audio dataset. The result obtained is shown in Figure 2. The depression group had a significantly smaller conversational network size compared with the healthy group. The conversational network size of the psychosis group was also smaller than the healthy group (though not significantly).
Per-scenario performance of the spurious

The everyday conversational network is a function of

Interpretation

The everyday conversational network is a function of

DISCUSSION

In this work, we proposed ECoNet, an everyday conversational network estimator, based on the unsupervised clustering of speech segments. ECoNet outperformed the existing audio-based speaker-counting system and was used to assess the relation of the conversational network size, a sociability measure, with mental health scores. Conversational network size was significantly correlated with mental health scores, such as depression severity.

ECoNet: Everyday Conversational Network Estimator

We demonstrated that a DNN-based VAD model could provide robust speech detection in free-living audio. Several VAD models, in comparison, gave a high error rate (see Table S1 in the supplementary material, available online) similar to the observations in Liaqat et al.’s work. The PyanNet-architecture-based DNN model had a lower error likely because of automatically learned bandpass filters for speech-discriminative features. We also showed that the spurious speaker detection can be framed as a machine learning problem, which gave a good detection performance (see Figure S1 in the supplementary material, available online). A multi-feature model likely better represents different spurious speaker-detection scenarios compared with a single-threshold approach in Xu et al.’s work. Per-scenario performance of the spurious speaker-detection model, e.g., in the presence of overlapping speech, background noise, etc., should be assessed for further improvements. Overall, ECoNet outperformed Crowd++ and its recent variations in the speaker-counting task (see Table S2 in the supplementary material, available online) possibly due to the robust VAD model, speaker-discriminative embeddings, and representative spurious speaker-detection model.

Conversational Network Size and Mental Health Parameters

The conversational network size, a sociability measure related to the (in-person) social network size, was significantly smaller in the depression group compared with the healthy group (see Figure 2). The average conversational network size had a correlation coefficient of −0.56 (p < 0.01) with PHQ score (see Figure 4). These two observations from our study fall in line with the social isolation hypothesis that implicates reduced social network size and lower sociability to depression. The conversational network size of the psychosis group was also smaller compared with the healthy group, but not significantly so, possibly reflecting a different mechanism of social isolation in psychotic disorders. In general, reduced conversational network size indicates reduced sociability and was, as expected, associated with the higher stress-related levels (higher PSS/ GAD scores), and lower positive health scores, (e.g., low DTS scores).

Correlation Between Conversational Network Size and Mental Health/Personality Scores

We computed the Pearson’s correlation coefficient between the average conversational network size of the participants and their mental health/personality scores. The result is shown in Figure 3. As an example, the correlation between the conversational network size and depression severity (PHQ score) was −0.56 (p < 0.01), shown in Figure 4. This association was not affected when accounted for possible confounding variables. The PHQ score had a coefficient of −0.10 (z-value: −3.14**) in predicting conversational network size within a generalized linear model with confounding variables included as predictors (see Section S-II-G in the supplementary material, available online).

**FIGURE 2.** Estimated conversational network size difference in the three population groups present in the clinical audio dataset.
Conversational network size estimated from ECoNet has potential (mental) health applications\(^1\)–\(^3,5\) as well as likely applications in crowd monitoring, interaction assessment in educational settings, etc. Estimation of fine-grained conversation features (turn-taking behaviors and engagement) using ECoNet could further enable new clinical and non-clinical applications.

Limitations and Future Work
Although ECoNet outperformed existing audio-based speaker-counting systems, it can be further improved. For example, a context-adaptive ECoNet that adapts the processing to detected scenarios, e.g., outdoor or indoor, could be explored. Each module within ECoNet could also be improved. The VAD model, for instance, could be extended with attention mechanisms for better temporal context. The ECoNet modules were developed and validated using diverse public datasets, such as the DIHARD dataset,\(^{17}\) consisting of audio from representative real-life scenarios. The clinical audio dataset from our pilot study, although a large audio dataset, could not be used for the development as the audio cannot be listened to, ensuring privacy as required in the study. A large labeled free-living audio dataset would be helpful to further improve the VAD models and ECoNet in general.

We designed ECoNet to be privacy-preserving as a speech-based system might raise privacy concerns. No analysis of the spoken contents is done, raw audio data need not be stored, and all processing can potentially be done in the individual's device or on a private network. We analyzed the recordings offline in this work, but the execution time was much smaller compared with the audio duration (see Section S-I-G). Thus, a (semi-)real-time processing on the target device/private network could be possible. We will investigate ECoNet’s deployment on target devices and continue to assess the privacy implications in future developments.

We investigated the potential of conversational network size obtained from the ECoNet in mental health applications. The depression severity, for example, was

\[ \text{CONVERSATIONAL NETWORK SIZE} \]

| Score          | 1.00*** | -0.18*** | -0.15*** | -0.12*** | 0.27 | 0.12 | 0.40* | -0.37*** | 0.20 | 0.47*** | 0.51*** | 0.52*** | 0.48*** | 0.20 |
|----------------|---------|-----------|-----------|-----------|------|------|-------|-----------|------|---------|---------|---------|---------|-------|------|
| PHQ score      | 0.16*** | 1.00***   | 0.81***   | 0.77***   | 0.20 | 0.30 | -0.20 | -0.37*** | 0.26 | -0.75*** | -0.54*** | -0.77*** | -0.79*** | -0.38*** |
| PSS score      | -0.58***| 0.81***   | 1.00***   | 0.77***   | -0.33 | -0.31 | 0.58*** | 0.69***   | -0.51 | -0.75*** | -0.73*** | -0.80*** | -0.80*** | -0.38*** |
| GAD score      | 0.52*** | 0.78***   | 0.76***   | 0.79***   | 0.29 | 0.30 | -0.26 | -0.21*** | 0.72*** | -0.76*** | -0.81*** | -0.81*** | -0.85*** | -0.65*** |
| Mini-IPIP       | 0.27 | -0.31 | -0.33 | -0.28 | 0.10*** | 0.41* | 0.23 | -0.20 | 0.56*** | 0.26 | 0.44* | 0.50 | 0.17 | 0.16 |
| Extraversion score | 0.12 | -0.36 | -0.36 | 0.41* | 1.00*** | 0.41* | -0.27 | 0.53*** | 0.40* | 0.48* | 0.32 | 0.30 | 0.35*** |
| Agreeableness score | 0.09 | -0.69*** | -0.69*** | -0.71*** | 0.33 | 0.41* | 1.00*** | -0.58*** | 0.21 | 0.50*** | 0.52*** | 0.52*** | 0.87*** | 0.38*** |
| Conscientiousness score | -0.53*** | 0.72*** | 0.69*** | 0.79*** | -0.29 | -0.27 | -0.84*** | 1.00*** | -0.29 | -0.62*** | -0.51*** | -0.70*** | -0.69*** | -0.34*** |
| Neuroticism score | 0.20 | -0.26 | -0.31 | -0.26 | 0.58*** | 0.52*** | 0.21 | -0.29 | 1.00*** | 0.29 | 0.44* | 0.22 | 0.22 | 0.19 |
| Mini-IPIP       | 0.47*** | -0.70*** | -0.75*** | -0.78*** | 0.29 | 0.40* | 0.62*** | 0.62*** | 0.29 | 1.00*** | 0.93*** | 0.95*** | 0.93*** | 0.82*** |
| Tolerance score | 0.51*** | -0.54*** | -0.73*** | -0.59*** | 0.44* | 0.46* | 0.50*** | -0.27*** | 0.44* | 0.50*** | 0.94*** | 0.94*** | 0.94*** | 0.79*** |
| Appraisal score | 0.50*** | -0.77*** | -0.81*** | -0.85*** | 0.26 | 0.30 | 0.52*** | -0.70*** | 0.22 | 0.49*** | 0.81*** | 0.71*** | 0.91*** | 0.69*** |
| Absorption score | 0.48*** | -0.79*** | -0.80*** | -0.79*** | 0.17 | 0.20 | 0.67*** | -0.69*** | 0.22 | 0.49*** | 0.81*** | 0.91*** | 0.91*** | 0.80*** |
| Regulation score | 0.20 | -0.29*** | -0.28*** | -0.45*** | 0.16 | 0.39* | 0.36* | -0.34 | 0.19 | 0.82*** | 0.72*** | 0.69*** | 0.80*** | 1.00*** |

**FIGURE 3.** Correlation between conversational network size and mental health/personality scores obtained with clinical questionnaires for participants in our clinical audio dataset.

**FIGURE 4.** Average conversational network size and PHQ score (correlation: 0.56, \( p < 0.01 \)) per participant, as assessed in the clinical audio dataset.
significantly correlated with conversational network size (see Figure 4). However, the correlation was lower and not significant when considering the individual groups (see Section S-II-E). Additional features derived from ECoNet might be helpful to better model depression severity differences within a group. It is to be noted that average daily conversational network size provides only one view of sociability, assigning importance to daily interactions even if with the same speakers across days. Other views of sociability, such as the total network size over a monitoring period, could also be relevant to study and will be explored in future work. As discussed in the “Interpretation” section, a limitation of ECoNet is that an individual’s involvement in the conversation cannot be ascertained. We will extend ECoNet with a conversational scene analysis module and explore other sociability measures that could be used to develop a mental health classification or severity prediction models in future work. Both the conversation properties of the most significant speakers, including the target individual, and characteristics of their utterances could be relevant to study (see Section S-II-H, available online). We will validate ECoNet’s capability to extract fine-grained sociability measures in future work. Measures related to interaction dynamics and turn-taking behaviors could help clarify fulfillment or loneliness, dissatisfaction, etc., resulting from social interactions, which are relevant in mental health applications.

CONCLUSION

We developed ECoNet, an audio-based automatic estimator of everyday conversation network from free-living audio, to complement the self-report-based assessments of sociability measures with automatically inferred measures. Several sociability measures could be extracted from the estimated conversation network, such as conversational network size, a task at which ECoNet outperformed the existing system. Conversational network size inferred using ECoNet was found to be significantly correlated with mental health (e.g., depression severity) in our pilot study on 32 participants with diverse mental health conditions. Our work demonstrates the relevance of an automatic sociability measure estimation from free-living audio for mental health applications. Further evaluation and validation with a larger population are required in future work to build on the observations from our work.

ACKNOWLEDGMENTS

This work involved human subjects or animals in its research. Approval of all ethical and experimental procedures and protocols was granted by Institutional Review Board (IRB), Rice University and Baylor College of Medicine and Affiliated Hospitals under Application no. IRB-FY2018-244 and H-41811, and performed in line with the United States Code of Federal Regulations at 45 CFR 46 and 21 CFR 56.)

REFERENCES

1. D. J. Vandervoort and V. B. Skorikov, “Physical health and social network characteristics as determinants of mental health across cultures,” Cur. Psychol., vol. 21, no. 1, pp. 50–67, 2002, doi: 10.1007/BF02903159.
2. S. Brinkhues et al., “Socially isolated individuals are more prone to have newly diagnosed and prevalent type 2 diabetes mellitus-the Maastricht study,” BMC Public Health, vol. 17, no. 1, pp. 1–12, 2017, doi: 10.1186/s12889-017-4948-6.
3. V. Gianfredi et al., “Association between social network characteristics and prevalent and incident depression: The Maastricht study,” J. Affect. Disord., vol. 293, pp. 338–346, 2021, doi: 10.1016/j.jad.2021.06.046.
4. B. Marroquin, V. Vine, and R. Morgan, “Mental health during the COVID-19 pandemic: Effects of stay-at-home policies, social distancing behavior, and social resources,” Psychiatry Res., vol. 293, 2020, Art. no. 113419, doi: 10.1016/j.psychres.2020.113419.
5. Z. I. Santini et al., “Social disconnectedness, perceived isolation, and symptoms of depression and anxiety among older Americans (NHISAP): A longitudinal mediation analysis,” Lancet Public Health, vol. 5, no. 1, pp. e62–e70, 2020, doi: 10.1016/S2468-2667(19)30230-0.
6. T. R. Insel, “Digital phenotyping: Technology for a new science of behavior,” JAMA, vol. 318, no. 13, pp. 1215–1216, 2017, doi: 10.1001/jama.2017.11295.
7. D. C. Mohr, M. Zhang, and S. M. Schueller, “Personal sensing: Understanding mental health using ubiquitous sensors and machine learning,” Annu. Rev. Clin. Psychol., vol. 13, pp. 23–47, 2017, doi: 10.1146/annurev-clinpsych-032816-044949.
8. M. Yamamoto et al., “Using speech recognition technology to investigate the association between timing-related speech features and depression severity,” PLoS One, vol. 15, no. 9, 2020, Art. no. e0238726.
9. M. R. Mehl, “The electronically activated recorder (EAR) a method for the naturalistic observation of daily social behavior,” Cur. Directions Psychol. Sci., vol. 26, no. 2, pp. 184–190, 2017, doi: 10.1177/0963721416680611.
10. T. Elmer and C. Stadtfeld, “Depressive symptoms are associated with social isolation in face-to-face interaction networks,” Sci. Rep., vol. 10, no. 1, 2020, Art. no. 1444, doi: 10.1038/s41598-020-58297-9.
11. C. Xu et al., “Crowd unsupervised speaker count with smartphones,” in Proc. ACM Int. Joint Conf. Pervasive Ubiquitous Comput., 2013, pp. 43–52, doi: 10.1145/2493432.2493435.
12. W. Wang, F. Seraj, N. Meratnia, and P. J. Havinga, “Speaker counting model based on transfer learning from SinCnet bottleneck layer,” in Proc. IEEE Int. Conf. Pervasive Comput. Commun., 2020, pp. 1–8, doi: 10.1109/PerCom45495.2020.9127390.

13. C. Xu, S. Li, Y. Zhang, E. Miluzzo, and Y.-F. Chen, “Crowdsensing the speaker count in the wild: Implications and applications,” IEEE Commun. Mag., vol. 52, no. 10, pp. 92–99, Oct. 2014, doi: 10.1109/MCOM.2014.6917408.

14. F.-R. Stötter, S. Chakrabarty, B. Edler, and E. A. Habets, “CountNet: Estimating the number of concurrent speakers using supervised learning,” IEEE/ACM Trans. Audio, Speech, Lang. Process., vol. 27, no. 2, pp. 268–282, Feb. 2019, doi: 10.1109/TASLP.2018.2877892.

15. D. Liaqat, R. Wu, A. Gershon, H. Alshaer, F. Rudzicz, and E. de Lara, “Challenges with real-world smartwatch based audio monitoring,” in Proc. 4th ACM Workshop Wearable Syst. Appl., 2018, pp. 54–59, doi: 10.1145/3211960.3211977.

16. H. Bredin et al., “Pyannote.audio: Neural building blocks for speaker diarization,” in Proc. ICASSP IEEE Int. Conf. Acoust., Speech Signal Process., 2020, pp. 7124–7128, Feb. 2019, doi: 10.1109/ICASSP40776.2020.9052974.

17. N. Ryant et al., “The third DIHARD diarization challenge,” 2020, arXiv:2012.01477.

18. D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, “X-vectors: Robust DNN embeddings for speaker recognition,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process., 2018, pp. 5329–5333, doi: 10.1109/ICASSP.2018.8461375.

19. J. Cao, “Objective sociability measures from multimodal smartphone data and unconstrained day-long audio streams,” Ph.D. dissertation, ECE, Rice Univ., Houston, TX, USA, 2019. [Online]. Available: https://hdl.handle.net/1911/105989

20. S. Saeb et al., “Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: An exploratory study,” J. Med. Internet Res., vol. 17, no. 7, 2015, Art. no. e175, doi: 10.2196/jmir.4273.

BISHAL LAMICHHANE is currently a graduate student with the Department of Electrical and Computer Engineering, Rice University, Houston, TX, USA. His research interests include machine learning applications in healthcare, with a current focus on mental health applications. Lamichhane received the M.S. degree in electrical engineering from Eindhoven University of Technology (TU/e), in 2014. He is the corresponding author of this article. Contact him at bishal.lamichhane@rice.edu.

NIDAL MOUKADDAM is currently an associate professor with the Baylor College of Medicine, Menninger Department of Psychiatry, Rice University, Houston, TX, USA. Her research interests include leveraging technology in diagnosing, tracking, and treating mental illness, and enhancing wellness levels. Moukaddam received the M.D. degree from American University of Beirut, in 1999. Contact her at nidalm@bcm.edu.

ANKIT B. PATEL is currently an assistant professor with the Baylor College of Medicine, Department of Neuroscience, Rice University, and an adjunct professor with the Department of Electrical and Computer Engineering and Computer Science and Department of Cancer Biology, MD Anderson Cancer Center, Houston, TX, USA. His research interests include (deep) machine learning, explainable artificial intelligence, and its application to high-impact biomedical/physical sciences. Patel received the Ph.D. degree in applied mathematics from Harvard University, in 2008. Contact him at Ankit.B.Patel@rice.edu.

ASHUTOSH SABHARWAL is currently a professor with the Department of Electrical and Computer Engineering, Rice University, Houston, TX, USA. His research interests include wireless communications and healthcare technologies, focusing on digital health. Sabharwal received the Ph.D. degree in electrical engineering from The Ohio State University, in 1999. He is a fellow of the IEEE and National Academy of Inventors. Contact him at ashu@rice.edu.