Digital Phenotyping for Detecting Depression Severity in a Large Payor-Provider System: Speech and Language Model Performance

Bradley Karlin, Doug Henry, Ryan Anderson, Salvatore Cieri, Mchael Aratow, Elizabeth Shriberg, Michelle Hoy

Table of Contents

Digital Phenotyping for Detecting Depression Severity in a Large Payor-Provider System: Speech and Language Model Performance

Bradley Karlin[1] PhD; Doug Henry[1] PhD; Ryan Anderson[1] PhD; Salvatore Cieri[1] LCSW; Mchael Aratow[2] MD; Elizabeth Shriberg[3] PhD; Michelle Hoy[3] LP

[1] Highmark Blue Cross Blue Shield Pittsburgh US
[2] Ellipsis Health San Francisco US

Corresponding Author:
Mchael Aratow MD

Abstract

Background: There is considerable need to improve and increase the detection and measurement of depression. The use of voice as a digital biomarker of depression represents a considerable opportunity for transforming and accelerating depression identification and treatment; however, research to date has primarily consisted of small-sample feasibility or pilot studies incorporating highly controlled applications and settings. There has been limited examination of the technology in real-world use contexts.

Objective: There is considerable need to improve and increase the detection and measurement of depression. The use of voice as a digital biomarker of depression represents a considerable opportunity for transforming and accelerating depression identification and treatment; however, research to date has primarily consisted of small-sample feasibility or pilot studies incorporating highly controlled applications and settings. There has been limited examination of the technology in real-world use contexts.

Methods: 2086 recordings of case management calls with verbally administered PHQ-9 surveys were analyzed using the ML model after the portions of the recordings with the PHQ-9 survey were manually redacted. The recordings were divided into a Development set (n=1336) and Blind set (n=671) and PHQ-8 scores were provided for the Development set for ML model refinement while PHQ-8 scores from the Blind set were withheld until after ML model depression severity output was reported.

Results: The Development set and Blind set were well matched for age, gender and depression severity, with mean and standard deviation of age of the Development set 53.7+/- 16.3 years and the Blind set 51.7 +/- 16.9 years, biological sex of the Development set 68.1% female and the Blind set 68.8% female and mean and standard deviation of the PHQ-8 scores of the Development set 10.5 +/- 6.1 and the Blind set 10.9 +/- 6.0 respectively. The Concordance Correlation Coefficient (CCC) for the test of the ML model on the Development set was pc=0.57 and for the Blind set pc=0.54, while the MAE for the Development set was 3.91 and for the Blind set was 4.06, demonstrating strong model performance. This performance was maintained when dividing each set into subgroups of age brackets (<=39, 40-64 and >=65), biological sex, and the four categories of Social Vulnerability Index (SVI, an index based on 16 social factors) with CCCs ranging from pc=0.44-0.61. Performance at PHQ-8 threshold score cutoffs of 5, 10, 15 and 20 representing the depression severity categories of none, mild, moderate, moderately severe and severe (>=20) respectively, expressed as Receiver Operating Characteristic Curve – Area Under the Curve (ROC-AUC) values, varied between 0.79 and 0.83 in both the Development and Blind sets.

Conclusions: Overall, the findings suggest that voice may have significant potential for detection and measurement of depression severity over a variety of ages, gender and socioeconomic categories that may enhance treatment, improve clinical decision-making, and enable truly personalized treatment recommendations.

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?
✔ **Please make my preprint PDF available to anyone at any time (recommended).**

Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.
Only make the preprint title and abstract visible.
No, I do not wish to publish my submitted manuscript as a preprint.

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✔ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain v

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in <a href="http

No. Please do not make my accepted manuscript PDF available to anyone.

Original Manuscript

Digital Phenotyping for Detecting Depression Severity in a Large Payor-Provider System: Speech and Language Model Performance

Bradley E Karlin[1,2], PhD, MSCP, MBA; Doug Henry[1], PhD; Ryan J Anderson[1], PhD; Salvatore Cieri[1], LCSW; Michael Aratow[3*], MD; Elizabeth Shriberg[3], PhD; Michelle Hoy[3], LPC

[1]Highmark Health, Pittsburgh, PA
[2]Johns Hopkins Bloomberg School of Public Health, Baltimore, MD
[3]Ellipsis Health, San Francisco, CA

Corresponding Author:
Michael Aratow, MD
548 Market Street, PMB 49051
San Francisco, CA 94104-5401
United States
Phone: 1 4159913552
Email: mike@ellipsishealth.com

Abstract

Background:
There is considerable need to improve and increase the detection and measurement of depression. The use of speech as a digital biomarker of depression represents a considerable opportunity for transforming and accelerating depression identification and treatment; however, research to date has primarily consisted of small-sample feasibility or pilot studies incorporating highly controlled applications and settings. There has been limited examination of the technology in real-world use contexts.

Objective:
The current study evaluated the performance of a machine learning (ML) model examining both semantic and acoustic properties of speech in predicting depression across more than 2,000 real-world interactions between health plan members and case managers.

Methods:
2086 recordings of case management calls with verbally administered PHQ-9 surveys were analyzed using the ML model after the portions of the recordings with the PHQ-9 survey were manually redacted. The recordings were divided into a Development set (n=1336) and Blind set (n=671) and PHQ-8 scores were provided for the Development set for ML model refinement while PHQ-8 scores from the Blind set were withheld until after ML model depression severity output was reported.

Results:

The Development set and Blind set were well matched for age, gender and depression severity, with mean and standard deviation of age of the Development set 53.7+/- 16.3 years and the Blind set 51.7 +/- 16.9 years, biological sex of the Development set 68.1% (910 out of 1,336 participants) female and the Blind set 68.9% (462 out of 671 participants) female and mean

and standard deviation of the PHQ-8 scores of the Development set 10.5 +/- 6.1 and the Blind set 10.9 +/- 6.0 respectively. The Concordance Correlation Coefficient (CCC) for the test of the ML model on the Development set was p_c=0.57 and for the Blind set p_c=0.54, while the MAE for the Development set was 3.91 and for the Blind set was 4.06, demonstrating strong model performance. This performance was maintained when dividing each set into subgroups of age brackets (<=39, 40-64 and >=65), biological sex, and the four categories of Social Vulnerability Index (SVI, an index based on 16 social factors) with CCCs ranging from p_c=0.44-0.61. Performance at PHQ-8 threshold score cutoffs of 5, 10, 15 and 20 representing the depression severity categories of none, mild, moderate, moderately severe and severe (>=20) respectively, expressed as Receiver Operating Characteristic Curve – Area Under the Curve (ROC-AUC) values, varied between 0.79 and 0.83 in both the Development and Blind sets.

Conclusions:

Overall, the findings suggest that speech may have significant potential for detection and measurement of depression severity over a variety of ages, gender and socioeconomic categories that may enhance treatment, improve clinical decision-making, and enable truly personalized treatment recommendations.

Keywords

depression; vocal biomarkers; artificial intelligence; behavioral health; machine learning; healthcare; case management; acoustic; semantic; digital phenotyping

Introduction

Background

The prevalence and impact of behavioral health (BH) problems are at an all-time high. As many as 1 in 3 individuals throughout the United States have a BH condition [1]. Rates of subclinical needs are even higher, fueled in part by the psychological and social effects of the pandemic; approximately 1 in 2 individuals report one or more symptoms of depression or anxiety [2]. In fact, the prevalence of depression symptoms increased more than three-fold during COVID-19 [3]. At the same time, only 40% of those with BH conditions, and even fewer with subclinical needs, receive care of any kind, due to challenges with and delays in detection, perceived need, stigma, a paucity of providers and other factors; and less than 15 percent of individuals with serious BH conditions receive minimally adequate treatment [4]. For those who do receive care, there is an average lag time of 11 years from the time of symptom onset to first treatment, during which time symptoms often worsen and other comorbidities may develop [4].

The current state of BH care and high levels of unmet need reflect a reactive and downstream approach to the identification and treatment of BH problems that has characterized the industry for decades. Effectively and efficiently meeting BH needs requires a more proactive, upstream, and personalized approach that meets individual needs earlier in their clinical trajectories with right-sized and person-fit interventions [5]. Emerging innovations in data science and technology, particularly developments in advanced data analytics and the availability of high-quality, patient driven digital interventions, present unprecedented opportunities to transform and innovate the field of BH care and reduce enduring, high rates of unmet need.

One particularly promising innovation for advancing detection and delivery of proactive, personalized, and data-informed treatment is digital phenotyping. Digital phenotyping involves the detection of phenotypes, or behavioral signals, that may indicate or predict the presence of a BH problem. Translated by machine learning (ML) models and collected through passive data collection via smart phones, wearables, or other communication devices, these data signals may serve as clinically and, potentially preclinically, useful markers of BH problems. The potential relevance and utility of digital phenotyping, which has been identified as the "next frontier" for personalized and proactive care within the field of oncology [6], has attracted particular interest and attention in the field of BH care and personalized psychiatry, with recent calls for accelerated applications to clinical practice [7-9]. In their review of research in this area, Huckvale et al [7] declared, "Many . . . studies appear to anticipate that digital phenotyping should play a role in routine clinical practice, for example by enhancing aspects of clinical diagnosis and treatment through earlier detection of condition onset, relapse or treatment response."

Most of the research examining digital phenotyping for the detection of BH problems has focused on detection of depression [10]. The opportunity to engage passive and objective ML technology for better detecting depression presents particular opportunities in light of the fact that depression is undetected in approximately 50% of individuals with the condition in high-income countries, and in 80-90% of individuals with depression in low- and middle-income countries [11]. In addition to opportunities that automated detection of depression provides

for increasing low screening rates in most clinical and community settings, ML presents significant promise for overcoming under-reporting and under-detection due to stigma, lack of evaluative service access, misattribution of symptoms to physical illness or age-related factors, or under-recognition of symptoms. In addition, the use of ML for detecting depression offers significant potential for increasing *earlier* identification and intervention, enhancing clinical efficiency through more accurate triage and treatment performance monitoring, improving fidelity through the use of objective measures, providing decision support, and personalizing BH care. As Galatzer-Levy and Onnela [12] recently declared, "Ultimately, the development of clinically meaningful digital measurements and their implementation in real-world contexts will permit optimized and personalized treatments targeted to the individual's emergent presentation and needs."

Prior Work

Among the most promising applications of digital phenotyping is the use of speech as a vocal biomarker of depression and other BH conditions [13]. The application of speech analysis in this context includes models for moment-by-moment analysis of the semantic patterns ("what" is said) and/or the acoustic properties (x e.g. tone, pitch, loudness, duration, articulation, transitions, prosody) of speech. Increasing research has demonstrated the promise of speech analysis, including generally increasing accuracy in overall detection of depression and other conditions [7]. Despite this promise, research to date has been primarily conducted in controlled contexts and uses, and there has been very limited examination or application of this technology in real-world settings [7,14]. As Koutsouleris et al [14] recently noted, "[W]hile these innovations promise to revolutionize health care, little progress has been made toward real precision mental health applications. Implementation of these applications is often an afterthought."

Research on the use of speech analysis for measuring BH symptoms has consisted primarily of small-sample feasibility or pilot studies with non-representative samples [7,14-16]. For example, in a scoping review of speech analysis for measuring mood disorders conducted by Flanagan et al [15], approximately 80% of studies were pilot or feasibility studies with sample sizes ranging from 1 to 73 participants. Similarly, in their review, Chia and Zhang [10] reported a mean sample size of 60. Moreover, many studies have consisted of analysis of "toy" data sets or controlled proof-of-concept studies involving highly controlled designs that, while promising for establishing the potential of a technological innovation, have yielded findings that are not necessarily generalizable or have utility or effectiveness for real-world use [10,17]. These designs include use of analog speech tasks (e.g., responding to a singular question, reading formulated passages, answering questions about everyday life, often referred to as "closed-form" tests) that are often not comparable to real-world clinical settings or real-life contexts.

In addition, many studies examining speech analysis in the BH context have had important methodological limitations, including frequent reporting of selected metrics, such as reporting of sensitivity without specificity, leading to the recent call for research in this area to report multiple metrics, including robust metrics, such as the concordance correlation coefficient (CCC), that are not as biased to specific context, use case, and data label distributions [13]. Many studies have also relied on binary classifications (above or below cut-off score for clinical significance) for screening tools, which limit opportunities for promoting precision and personalization in BH care. Further, research on speech analysis in detecting and measuring BH symptoms has almost exclusively relied on the use of single methods of analysis

(predominantly acoustic analysis). Opportunities for leveraging and combining analysis of acoustic and semantic properties of speech may yield greater accuracy and precision in detecting and predicting BH conditions.

Goal of this Study

As the foregoing reveals, the application of digital phenotyping within BH care has approached a defining moment and key turning point for the field. In their review of the current state of digital phenotyping within the field of BH, Huckvale et al [7] have urged for "practical and coordinated action . . . to help accelerate both research and the ultimate development of real-world health applications for digital phenotyping". In an effort to help advance real-world application of digital phenotyping for promoting earlier and automated detection and measurement of depression, the current study was conducted evaluating the performance of a ML model of the semantic and acoustic properties of spoken language in predicting depression in a naturalistic context by analyzing more than 2,000 interactions between health plan members and case managers. Additionally, the study sought to test model performance beyond "presence/absence" dichotomous predictions, examining classificatory accuracy at multiple levels of depression from none or minimal to severe. Furthermore, model performance was tested across age, sex, and sociodemographic factors and in BH and non-BH case management contexts. This project, which is unique in its breadth and scope, aims to assess the accuracy of speech analysis for detecting and measuring depression severity in routine clinical settings. We hypothesize that the machine learning models used in this study will demonstrate robust predictive accuracy across variations in age, gender, care management context, and Social Vulnerability Index.

Methods

Experimental Design

The current quality improvement project evaluated the performance of the combined semantic-acoustic ML speech analysis model in predicting depression severity from existing recordings of case management calls, with BH case managers who are licensed independent mental health providers. Specifically, the performance of the ML model in care management conversations between insured members and BH case managers was evaluated by retrospectively comparing the actual scores from the PHQ-9 administered by the case managers. The predicted PHQ-8 scores were derived from the qualities (acoustic biomarkers and semantic content) of vocal productions of the same members conversing with care managers while engaged in discussion other than the PHQ-8 administration. It was secondarily sought to examine model performance in non-BH contexts where the PHQ-9 is not routinely administered using a subsample of calls with non-BH case managers. For both BH and non-BH calls, model predictions were compared to PHQ-8 scores from an associated metadata file.

On each of the calls analyzed, the PHQ-9 was verbally administered. Members consented to the recording of the call for quality and training purposes. This study was designated as a quality improvement project by the Institutional Review Board (IRB) of the Allegheny Health Network and therefore exempt from ongoing IRB oversight. The project was also reviewed and approved by the Highmark Health Enterprise Data Governance Committee to ensure that it comported with internal data protection standards and applicable privacy, legal and regulatory

requirements.

Measures

Depression severity

The Patient Health Questionnaire (PHQ-9) is a widely-utilized self-report measure of depression symptom severity. Frequency of depression symptoms are endorsed by patients using a four-point scale, ranging from 0 ("Not at all") to 3 ("Nearly every day"). PHQ-9 scores range from 0 to 27. Higher scores reflect greater depression severity. Scores of 0 to 4 are classified as "none to minimal," 5 to 9 are classified as "mild," 10 to 14 are classified as "moderate," 15 to 19 are classified as "moderately severe," and 20 to 27 are classified as "severe." The PHQ-9 has been shown to be an internally consistent, valid, and reliable measure of depression severity [18,19]. For the present study, the last item of the PHQ-9, which assesses for suicidal or self-injurious thoughts, was omitted given the use of archival data where further probing of responses was not feasible. Its inclusion requires different clinical considerations and handling in research settings. The adapted scale with item nine removed is referred to as the PHQ-8 and has been shown to have strong psychometric characteristics, including the ability to accurately predict depression [20].

ML Speech Analysis Model

The semantic-acoustic model evaluated in this study has demonstrated robust results for accurate prediction of depression symptom severity and acceptable rates of error [21-23]. The proprietary ML model includes both acoustic and semantic models.

The acoustic model takes as input the raw speech signal (rather than precomputed features such as pitch or energy). The production acoustic workflow is built on a pretrained open-source wav2vec2 architecture [24] and is trained on proprietary audio data. The system consists of four segment models, each trained with specific configurations, and three segment fusion models that integrate outputs from the segment networks. Predictions from the segment fusion models are weighted to generate the final acoustic score.

The semantic model (referred to also as a natural language processing (NLP) model) takes as input the output of a commercial automatic speech recognition (ASR) system. The model is based on the Longformer architecture [25], designed to efficiently handle long conversational contexts using advanced mechanisms such as dilated sliding window attention. Model training involves a proprietary fine-tuning approach using depression-specific data, utilizing high-quality proprietary transcripts paired with PHQ scores. Further refinement is conducted using conversational samples, also labeled with PHQ scores.

Labeled training data comes from a large corpus of proprietary spoken language data sets labeled with PHQ-8 values. Both models take advantage of publicly available data for model pretraining, including text corpora for the NLP model.

To generate the final depression severity prediction, the outputs of the acoustic and language model are combined using a linear weighting; the weight is optimized using the CCC metric on the Development set Figure 1 illustrates the overall machine learning analysis, from data

preprocessing to prediction generation.

Figure 1. Deep learning architecture and processing pipeline.
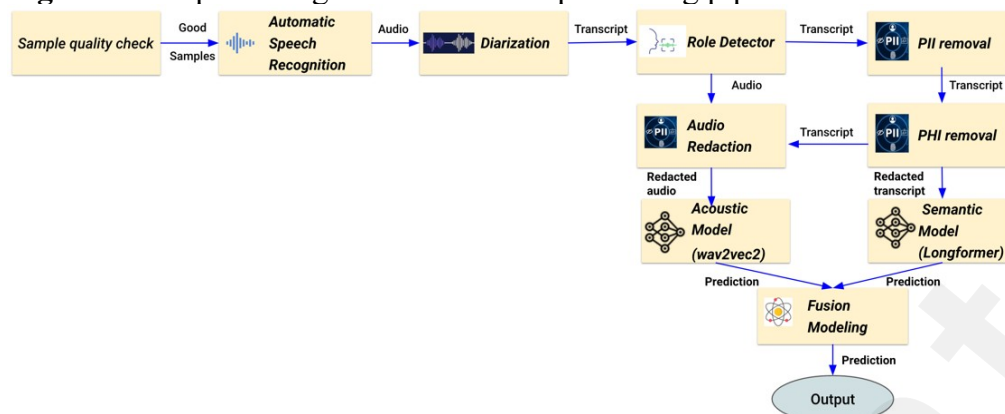
Identification of Case Management Calls and Metadata

A total of 2,626 recordings of case management calls were included. They took place between January 2019 and January 2023. Of these calls, 2,083 had full item-level data for the first eight items of the PHQ-9, which was collected verbally during the course of calls. Calls corresponded to unique members from 44 different U.S. states and were completed by 46 case managers. The majority of case managers completed multiple calls, and approximately one-third completed 20 or more calls. Each call recording had an associated metadata file containing member age, biological sex, zip code, whether the call was conducted by a BH case manager or non-BH (e.g., medical, surgical) case manager, and PHQ-9 item-level data. Exclusion criteria for recordings included member age less than 18 years, speechmail messages, presence of any speakers beyond the case manager and the member, recordings in which the member was not present, and diarization failures (failures to correctly segment audio into single-speaker time regions). These exclusion criteria constituted 76 of the 2,083 calls, leaving a total of 2,007 calls for the analysis.

Partitioning the Data

The evaluation was conducted in two phases. To establish datasets for both phases of the project, the 2,007 recordings and their metadata were partitioned by randomly assigning them to a development dataset (Dev Set) consisting of approximately two thirds of the total available calls (*n* = 1,336) and a test dataset (Blind Set) consisting of approximately one third of the total available calls (*n* = 671). There was no speaker overlap across these datasets. The partitions were constructed to ensure reasonably equal representation of the metadata, including the distribution of PHQ-9 severity (none, mild, moderate, moderately severe, severe). The Dev Set and Blind Set were securely delivered via secure file transfer protocol (SFTP) to Ellipsis Health, which performed all further data processing and analyses of the calls and metadata. The Blind Set was held back until phase one was completed.

The two datasets (Dev, Blind) were well matched; however, there were data curation errors such as inclusion of voicemails, conversations in a different language (mostly involving a language interpreter), and minors (<18 years old). Subsequent to the delivery of each dataset, 76 recordings, 44 from the Dev Set and 32 from the Blind Set, were found to meet exclusionary criteria through review of metadata (i.e., age) and through diarization tool flags indicating a single speaker or more than two speakers. These 76 recordings were removed from the analyses. However, because the audio tracks were not reviewed by the annotators, other

recordings meeting exclusionary criteria were included in the Dev and Blind Sets. An analysis was conducted, and it is estimated that inclusion errors constituted 3% of the total recordings analyzed.

Data Preprocessing

Upon receiving call recordings, Ellipsis Health performed diarization, automatic speech recognition and redaction of personally identifiable information (PII) using Amazon Web Service (AWS) Amazon Transcribe [26]. Redaction of Protected Health Information (PHI) was performed using Amazon Comprehend Medical (ACM) [27]. Speaker role detection and time stamp generation on turns in conversation (i.e., transition from member to case manager and vice versa) were performed using proprietary algorithms from Ellipsis Health. The redacted output of the automatic speech recognition process, which included transcripts of both the member and case manager with PHQ9 content removed, was used by the semantic model. Meanwhile, the diarized, redacted audio containing only the member's speech, with PHQ9 content masked, was used by the acoustic model. The outputs of the semantic and acoustic models, were used (after weighting) to arrive at the fused output, using the Dev dataset.

Manual Annotation

To remove the verbally administered PHQ-9 from the calls, a manual annotation process was performed to identify the regions in the transcripts where the PHQ-9 was administered. Ellipsis Health used a team of professionals separate from the team conducting tests of the ML model (i.e., ML team) to perform manual annotation, which consisted of annotators being presented with the case manager portion of the transcript from each call and having them identify the regions that contained the PHQ questions. These annotated regions of audio samples were masked in white noise for the acoustic model analysis, and the corresponding text was removed from the transcripts for the semantic model analysis.

Model Refinement on the Dev Set

In phase one, the semantic-acoustic model was applied to the Dev Set (n = 1,336), and hyperparameters (e.g., learning rate in optimization algorithms, number of hidden layers and number of iterations in training a neural network) were optimized to minimize the concordance correlation coefficient (CCC) [28].

Tests on the Blind Set

Phase two of the project was conducted to evaluate the performance of the semantic-acoustic ML model established in phase one. The Blind Set used in phase two was provided to Ellipsis Health without any accompanying PHQ-9 scores to ensure a blinded test of the model. Other than absence of PHQ-9 scores, the provided metadata categories were the same categories as provided for the Dev Set. Automatic speech recognition was conducted, followed by PII and PHI redaction of both the audio and transcript, using the same process as for the Dev set. Manual annotation of the recordings was performed as in phase one and the verbally administered PHQ-9 was masked in the audio file and removed from the call transcript. The ML team then conducted tests of the ML model to predict depression severity scores for the Blind Set and across the metadata subgroups of the Blind Set. Recorded PHQ-9 scores from the original calls in the Blind Set were subsequently provided to the ML team, PHQ-8 scores were then derived from these PHQ-9 scores and then compared to ML model predictions of the PHQ-8 scores to

complete the test of model accuracy.

In light of the fact that over- and underreporting are well-known phenomena of surveys, including on sensitive measures like the PHQ-9 [28,32,33], a preliminary exploration of the possible presence of such when responding to the PHQ-9 was conducted by examining for sizeable discrepancies between PHQ-8 labels and predicted depression scores. Over- and under-reporting was defined as a difference of ≥2 categories of classification between the model prediction and the PHQ-8 score, as this would likely cause a significant change in a care pathway for a patient, and this condition was found in 42 of the 2007 total recordings. Five licensed therapists were recruited to listen and rate the member for severity of depression symptoms (none, mild, moderate, severe). They were assigned recordings such that 1 therapist listened to each of the 42 recordings, but in 25 of those calls at least 2 therapists provided an additional assessment. The therapists were blinded to all information, including the PHQ score and section of the recording where the survey was administered, the model predictions, and demographic information. A PHQ-8 score predicted by ML model was defined as agreeing with a mental health provider assessment if their assessment was the same or within 1 severity category difference.

Metrics

The ML model results included CCC, mean absolute error (MAE), area under the receiver operating curve (ROC-AUC), and sensitivity and specificity at the point of equal error (EER) for the Dev Set and Blind Set. All classification analyses were conducted with the PHQ-8 as the criterion or observed score. Predicted scores from the ML regression models were binned according to the following PHQ-8 depression severity classifications: none or minimal (0 - 4), mild (5 - 9), moderate (10 - 14), moderately severe (15 - 19), and severe (20 - 24). Next, ROC analyses were conducted, comparing predicted to observed scores across five binary classifications at the four PHQ-8 cut-offs (5, 10, 15, 20): 0 to 4 versus 5 to 24, 0 to 9 versus 10 to 24, 0 to 14 versus 15 to 24, 0 to 19 versus 20 to 24. ROC-AUC and sensitivity and specificity at the rate of equal error were calculated at each cut-off and reported for the ML model.

Results

Comparison of Member-Level Metadata

Data demographic distributions for members were comparable across both datasets. Ages ranged from 18 to 98 years in the Dev Set to 18 to 92 years in the Blind Set (Table 1). Approximately two-thirds of members across the Dev Set (910 out of 1,336 participants, or 68.1%) and Blind Set (462 out of 671 participants, or 68.9%) were female. Member zip code was used to establish the Social Vulnerability Index (SVI) [29], which is based on 16 social factors, including socioeconomic status (e.g., below poverty, unemployed), household characteristics (e.g., single parent, ≥ 65 years), and housing type/transportation (e.g., crowding, no vehicle). In each dataset, members were predominantly in the low-moderate range for social vulnerability and the majority of calls were BH case management calls (Table 1).

Table 1. Distribution of metadata for Dev and Blind Sets.

Metadata	Dev Set (n = 1,336)	Blind Set (n = 671)
Age in years	M^a = 53.7, SD^b = 16.3, *Range* = 18 - 98	M = 51.7, SD = 16.9, *Range* = 18 - 92
Age (years), n (%)		
≤ 39	296 (22.2)	183 (27.3)
40 - 64	704 (52.7)	344 (51.3)
≥ 65	336 (25.1)	144 (21.4)
Biological sex, n (%)		
Female	910 (68.1)	462 (68.9)
Male	426 (31.9)	207 (30.8)
Undefined	0 (0)	2 (0.3)
SVI^c, n (%)		
1	279 (20.9)	150 (15.6)
2	617 (46.2)	306 (45.6)
3	335 (25.1)	162 (24.1)
4	102 (7.6)	51 (7.6)
Missing	3 (2.2)	2 (0.3)
Type of CM^d, n (%)		
BH^e	1,087 (81.4)	561 (83.6)
Non-BH	249 (18.6)	110 (16.4)
PHQ-8	M = 10.5, SD = 6.1, *Range* = 0 - 24	M = 10.9, SD = 6.0, *Range* = 0 - 24
PHQ-8^f, n (%)		
None or minimal	249 (18.6)	113 (16.8)
Mild	384 (28.7)	179 (26.7)
Moderate	328 (25.6)	174 (25.9)
Moderately severe	263 (19.7)	141 (21.0)
Severe	112 (8.4)	64 (9.5)

^{a}M: mean

^{b}SD: standard deviation

^{c}SVI: Social Vulnerability Index (1 = least vulnerable, 4 = most vulnerable).

Regression Results for Overall Tests of the ML Model

Results for the test of the ML model on the Dev Set (n = 1,336) produced a CCC of ρ_c = 0.57, which is superior to results expected by chance (ρ_c = 0.10 to 0.20). CCC showed minimal decrease in the test of the ML model on the Blind Set (ρ_c = 0.54, n = 671). Furthermore, MAE for the ML model tests across datasets were 3.91 and 4.06 for the Dev Set and Blind Set, respectively. These values for MAE are equivalent to less than the score range (five points) of a single PHQ-8 severity classification.

Classification Results for Overall Tests of the ML Model

The ROC-AUC at PHQ-8 cut-off of 10 (i.e., "moderate" depression, the traditional cutoff for the majority of clinical care pathways [29,30]) was consistent for the ML model as applied to the Dev Set (0.83) and Blind Set (0.81), which are identical to the respective mean ROC-AUC values over the different cutoff points (see Table 2, top panel). In particular, results for the ML model (ROC-AUC = 0.81) on the Blind Set indicates the robustness of the model in its ability to identify individuals with PHQ-8 scores above 10 using novel call data (i.e., data without PHQ-8 labels and not previously used for model refinement).

Table 2. Regression and classification results for overall model tests.

	ML Model	
Statistic	Dev Set (n = 1,336)	Blind Set (n = 671)
CCC[a]	0.57	0.54
MAE[b]	3.91	4.06
ROC-AUC[c]		
M[d] across cut-offs[e]	0.83	0.81
cut-off 5	0.81	0.85
cut-off 10	0.83	0.81
cut-off 15	0.83	0.79
cut-off 20	0.83	0.79
Sens=Spec[f]		
M across cut-offs	0.74	0.73
cut-off 5	0.73	0.76
cut-off 10	0.75	0.72
cut-off 15	0.74	0.72
cut-off 20	0.76	0.72

[a]CCC: concordance correlation coefficient.
[b]MAE: mean absolute error.
[c]ROC-AUC: area under the receiver operating characteristic curve.
[d]M: mean across four cut-offs (5, 10, 15, 20).
[e]Cut-off numbers were chosen as these are the points where the depression severity category

boundaries occur.
[f]Sens=Spec: value of both sensitivity and specificity at point of Equal Error.

ROC-AUC values across the four cut-off thresholds and across both datasets ranged from 0.79 to 0.85 (Table 2). The lowest ROC-AUCs were observed for the ML model (ROC-AUC = 0.79) on the Blind Set at PHQ-8 cut-offs of 15 and 20 ("moderately severe" and "severe" depression). Of note, the size of the subsample of members in the Blind Set with scores ≥ 20 was only 64 members and may have contributed to the lower ROC-AUC values for this classification. See Table 1 for information on sample sizes across classifications for all three datasets.

As shown in Table 2, the mean sensitivity and specificity at the point of equal error across the four classifications was stable for ML model performance. Across the four PHQ-8 cut-off scores, sensitivity/specificity values ranged from 0.72 at a cut-off of 10, 15 and 20 for the ML model test on the Blind Set to 0.76 for the ML model test on the Dev Set at cut-off of 20 and on the Blind Set at cutoff of 5. As observed with ROC-AUC, values at the lower end of the range for sensitivity/specificity may have been affected by smaller subsample sizes (e.g., Blind Set with PHQ-8 ≥ 20).

Model Performance Across Metadata Subgroups

ML model performance was evaluated across metadata subgroups based on age in years (18 - 39, 40 - 64, ≥ 65), sex (male, female), and BH case management versus non-BH case management and across the four SVI levels (1 = least vulnerable, 4 = most vulnerable).
 As shown in Table 3, ML model performance between the two datasets (Dev Set and Blind Set) within their subgroups (Age, Sex, Type of CM Call, SVI) reveals both consistent and relatively similar ROC-AUC cut-off at 10 and EER values with ROC-AUC cut-off at 10 ranging from 0.81-0.83 and sensitivity/specificity at point of Equal error ranging from 0.73-0.75, implying good model stability and robustness. See Figures S1-S12 in Multimedia Appendix 1 for ROC curves (overall and per subgroup on Blind Set). CCC ranged from 0.44-0.61, with the lowest (0.44) occurring in the most highly socially vulnerable group in the Blind Set and the highest (0.61) occurring in both the least socially vulnerable group of the Dev Set and the ≥65 years age group in the Blind Set. In most cases, the lower CCC values occurred where sample sizes were approximately 100 or fewer individuals, our previous work [31] suggests a minimum count of approximately 200 individuals for robust estimates of prediction performance. MAE values ranged from 3.62 in the ≥65 years age group of the Blind Set to 4.57 in the non-BH group of the Blind Set. The two highest MAE values were associated with subgroups with sample sizes of approximately 100 or fewer participants, comparable to results for the CCC.

Table 3. Regression and classification metrics for model tests across metadata subgroups.

	ML Model							
Variable/Statistic	Dev Set (n = 1,336)			Blind Set (n = 671)				
Age in years (n)	≤ 39 (296)	40 - 64 (704)	≥ 65 (336)	≤ 39 (183)	40 - 64 (344)	≥ 65 (144)		
CCC[a]	0.58	0.55	0.58	0.57	0.47	0.61		
MAE[b]	3.91	4.00	3.77	3.93	4.32	3.62		
ROC-AUC[c] cut-off[d] 10	0.83	0.83	0.83	0.81	0.81	0.81		
Sens=Spec[e] cut-off 10	0.76	0.75	0.75	0.72	0.72	0.72		
Sex (n)	F[g] (910)		M[f] (426)		F (462)		M (207)	
CCC	0.56		0.58		0.53		0.54	
MAE	3.86		4.04		3.95		4.29	
ROC-AUC cut-off 10	0.83		0.83		0.81		0.81	
Sens=Spec cut-off 10	0.75		0.75		0.72		0.72	
Type of CM[h] Call (n)	BH[i] (1087)		Non-BH (249)		BH (561)		Non-BH (110)	
CCC	0.57		0.58		0.55		0.46	
MAE	3.92		3.86		4.00		4.40	
ROC-AUC cut-off 10	0.83		0.83		0.81		0.81	
Sens=Spec cut-off 10	0.75		0.75		0.72		0.72	
SVI[j]	1 (279)	2 (617)	3 (335)	4 (102)	1 (150)	2 (306)	3 (162)	4 (51)
CCC	0.61	0.57	0.57	0.46	0.61	0.54	0.47	0.44
MAE	3.76	3.90	3.98	4.24	3.83	4.04	4.18	4.57
ROC-AUC cut-off 10	0.83	0.83	0.83	0.83	0.81	0.81	0.81	0.81
Sens=Spec cut-off 10	0.75	0.75	0.75	0.75	0.72	0.72	0.72	0.72

[a]CCC: concordance correlation coefficient.

[b]MAE: mean absolute error.

[c]ROC-AUC: area under the receiver operating characteristic curve.

13

Lastly, within the Dev Set, 42 of the 1,336 recordings (3.1%) showed sizable discrepancies (divergence equivalent to two or more PHQ-8 categories) between administered PHQ-8 and ML-predicted depression score that could imply PHQ-8 response under-reporting (actual depression score much lower than predicted depression score) or over-reporting (actual depression score much higher than predicted depression score). A review of these discrepancies by five independently licensed clinicians, who were blinded to the administered score, yielded PHQ-8 categorization of members' vocalizations that were consistent with the ML model categorization twice as often as they were with the administered PHQ-8 score.

Discussion

Principal Results

speech The current evaluation, leveraging speech analysis to detect depression symptoms across different levels of severity within a large real-world clinical case management context, represents, to our knowledge, the first evaluation of its kind. Overall, the findings for the combined acoustic-semantic ML model demonstrate strong model performance across a variety of key metrics. The ROC-AUC values of approximately 0.81, the overall CCC value of 0.54 and mean of the sensitivity/specificity at the EER of 0.73 in the Blind Set demonstrate robust clinical support for the model's ability to accurately predict severity of depression. These results compare favorably to previous research, which has primarily relied on much smaller samples and incorporated pilot, simulation, or controlled study designs[32,33]. Whereas many prior studies have focused on application of acoustic or semantic speech analysis, this study reports on an approach that combined information from both semantic and acoustic based models. Future development and testing of speech analysis models should continue to explore the benefits of combining acoustic and semantic models.

It is noteworthy that the model performed consistently across all major subgroups and PHQ-8 classification levels, with ROC-AUCs values ranging from 0.79 to 0.85 and CCC values ranging from 0.44 to 0.61. However, the model did under-characterize individuals at the highest PHQ-8 level (i.e., "severe" depression), likely due to smaller sample size in this category of depression on which this model was trained. Among the most promising findings from the subgroup analyses was the model's strong accuracy in predicting

14

depression severity among older adults across key metrics. This finding is particularly significant given that older adults have the highest rates of undetected depression [34], are often less likely to recognize or self-report symptoms of depression [35], and may experience depression with fewer dysphoric symptoms and more somatic complaints, which can be misattributed to physical illness [36,37].

The model's ability to detect depression symptoms at lower severity levels offers significant real-world potential for early identification and person-fit and right-sized interventions earlier in individuals' clinical trajectories. In addition to its ability to classify depression presence (PHQ-8 <10 vs. ≥10), the model performed well across specific PHQ-8 severity levels, particularly in the minimal and mild ranges. This suggests promising applications for early, proactive, cost-effective, and lower-intensity interventions (e.g., digital interventions, BH coaching, peer/social support) that may prevent symptom progression and reduce relapse rates. Notably, beyond the personal and clinical benefits, earlier interventions and prevention of depression may also have significant financial implications, potentially reducing healthcare costs associated with advanced-stage depression treatment.

Beyond BH contexts, the model's performance in general medical (non-BH) case management calls suggests even greater potential for broadening depression detection. The model achieved a comparable level of performance (ROC-AUC cut-off 10= 0.81, Range = 0.79 to 0.85) in non-BH case manager calls, indicating potential for integrating depression detection into clinical decision-making in settings where the PHQ-9 is not routinely administered and where depression is often undetected [38]. Further, the potential financial impact of enhanced depression detection in non-BH contexts is considerable, especially given that the total cost of care for members with both a BH condition and a chronic physical health condition, experienced by many members in medical-surgical case management, is approximately three times higher than for those with the same condition but no BH diagnosis [39,40]. The foregoing notwithstanding, findings related to model performance in non-BH case management should be considered preliminary in light of the smaller sample size (*n* = 110, Blind Set sample). Accordingly, additional application of speech analysis in non-BH contexts and other physical health settings is warranted.

The findings with respect to the over- and under-reporting on the PHQ-8 offer insights into the potential for speech-based analysis to enhance depression detection objectivity [41-43]. Specifically, speech analysis may be less susceptible to bias, whether conscious or unconscious, relative to subjective report or traditional measurement. The observed trend of likely under-reporting on the PHQ-8 (i.e. the administered score being much lower than the predicted score) aligns with prior research on the impact of stigma, lower BH literacy, and other factors that contribute to self-report bias and under-reporting of depression [44]. On the other hand, over-reporting on the PHQ-9 (i.e., the administered score being much higher than the predicted score) may suggest heightened or exaggerated response tendencies, personality characteristics, or efforts to obtain help [45].

Finally, this study highlights important opportunities for speech analysis and other digital

15

phenotyping approaches to improve administrative and clinical workflows. It is noteworthy that case managers spent nearly 20% of total call time administering the PHQ-9. From an efficiency standpoint, this is time that could be better allocated to establishing a therapeutic alliance, collaborating to identify and define BH goals, addressing ambivalence and other potential obstacles to achieving those goals, and directly addressing the member's chief concerns. Greater efficiency also establishes opportunity for case managers to interact with more members. From a clinical process standpoint, time spent administering measures like the PHQ-9 can be awkward and even frustrating for members and may adversely affect rapport and engagement. Furthermore, inconsistencies in PHQ-9 administration can introduce errors or variability in measurement, potentially leading to misinterpretation of symptoms. In contrast, having objective, real-time data on depression severity could provide valuable insights for clinical decision-making and for providing proactive and personalized treatment plan.

Limitations

This study has several key strengths, including its large sample size, evaluation of speech analysis in a real-world clinical context, use of naturalistic conversations vs. analog speech tasks (such as reading defined passages or phrases, or repeating specified sounds), integration of both semantic and acoustic properties of speech, and analysis of model performance across subgroups and depression severity levels using numerous evaluation metrics. At the same time, there are several limitations that should be considered when interpreting the current findings and guiding future research

While the large sample size included in this real-world evaluation is unique in the field of digital phenotyping [15], the sample sizes for some of the sub-analyses, as noted above, including the analyses of the non-BH calls and the highest PHQ-9 severity level ("Severe") condition, were relatively small. Given this, caution should be exercised when interpreting these findings. Moreover, data on member race/ethnicity were not available. As such, the generalizability of the current results to different ethnic, cultural, and linguistic groups cannot be definitively determined. That said, the acoustic-semantic speech model was developed and trained on a very large and diverse sample [21,46,].

Further, prediction of depression by the model, like with any ML model, includes a degree of error or imprecision. In the current analysis, this was equivalent to approximately 4.06 points on average (the reported MAE) on the PHQ-9, which itself has imperfect accuracy [47]. As such, predicted scores should be interpreted with this in mind. With additional data, precision is likely to further increase.

In addition, the collection of the case management recordings on a single audio channel posed a challenge for this study, necessitating the use of automatic speech recognition (ASR) for conversion of speech-to-text), diarization for speaker separation, and speaker attribution labeling. While these processes generally have low error rates, they are not entirely error-free. Diarization was performed using a leading commercial tool as manual

16

processing of calls would have required listening to and annotating thousands of hours of recordings, a time-intensive process that is also prone to errors. Additionally, the use of automatic speech processing better reflects how an actual implementation would be performed in a real-world setting. However, diarization errors (i.e., poor or missing speaker separation) were encountered, and these errors propagated through the preprocessing and annotation pipeline, affecting both automated speaker attribution and removal of PHQ-8 content.

Furthermore, data curation errors observed (e.g., inclusion of voicemails, conversations with minors), inevitable in a real-world dataset of this kind, may have impacted performance (in both positive and negative directions); on balance, however, these likely did not have more than a negligible effect on the reported performance metrics. Many of these challenges and resulting errors may be attenuated in prospective implementations (versus the current retrospective-focused analysis) in the future given that: 1) current call management systems routinely record speakers on separate channels, significantly mitigating the challenges of diarization (many legacy systems do not have diarization capability, but some may be specially configured to do so), and 2) formal implementation of this technology within the case manager's workflow would limit inclusion of irrelevant (e.g., voicemail messages) or inappropriate (e.g. minors, different languages) calls, either through call management software technology or manual exclusion by the case manager according to inclusion/exclusion criteria.

Future studies and real-world deployments

The successful deployment of AI-driven speech analysis for depression detection requires careful integration into existing healthcare workflows. One promising approach is its integration into telehealth platforms, where it could facilitate real-time assessment and early detection during virtual consultations. Embedding the model into electronic health records (EHRs), virtual scribe technology or clinical decision-support systems could further enhance its utility by providing clinicians with objective, data-driven insights alongside traditional assessments. This study represents a step toward the rigorous validation of AI-based healthcare tools, ensuring their accuracy and reliability across diverse populations. For successful deployment, ensuring security, safety, and compliance with HIPAA, GDPR, and other data protection regulations is essential, along with continuous monitoring of system performance on test datasets to maintain reliability and accuracy. Additionally, clinician adoption depends on ensuring that the tool is user-friendly and seamlessly integrates into existing workflows without adding unnecessary burden.

Although the data used in this study are de-identified, future studies and real-world deployments should incorporate a protocol for obtaining explicit informed consent from members, ensuring ethical transparency and alignment with established guidelines for digital health interventions. One of the primary challenges in ML-based depression detection is the mitigation of bias, as algorithmic outputs may be influenced by dataset imbalances or systemic biases. In this study, bias evaluation was conducted across key

17

demographic subgroups, but future research should expand on bias mitigation strategies and assess ethical AI deployment frameworks to ensure equitable model performance across diverse populations.

Additionally, AI governance is a critical factor in real-world deployment, necessitating adherence to key principles such as transparency, fairness, and accountability. Transparency ensures that AI models operate in a manner that is understandable, explainable, and accessible to stakeholders, including clinicians and patients. Fairness requires that models are developed and validated in a way that minimizes bias and ensures equitable performance across diverse populations, preventing disparities in mental health assessments. Accountability involves establishing clear oversight mechanisms to monitor AI decision-making, ensuring that these technologies align with ethical standards, regulatory requirements, and best practices for patient care. Future research and implementation should prioritize these principles to foster trust and reliability in AI-driven mental health tools.

Finally, while the ML model demonstrated strong predictive performance, it is important to emphasize that this tool is intended for initial assessment and triage, not for medical diagnosis. The model is designed to support early identification and risk stratification, which should be followed by the clinician evaluation and judgment. This tool is not intended to replace traditional diagnostic methods.

Conclusion

There is an urgent need to enhance detection and measurement of depression. Implementing digital phenotyping through the use of speech as a digital biomarker of depression offers significant promise for improving and accelerating depression identification and treatment. In short, the current evaluation, involving the examination of combined acoustic and semantic speech analysis for predicting depression symptom severity across PHQ-9 classification levels in a large real-world clinical context, represents the first evaluation of its type. The results reported herein provide strong support for the application and utility of a readily available and unobtrusive biomarker, namely, what and how of spoken language, for detecting and measuring depression in real-world practice at this important time. This easily accessible biomarker has significant potential for application in health care settings, ranging from "pre-clinical" case management contexts to patient-provider interactions. It is hoped that the current findings help to advance the development and application of novel ML technologies for automating and enhancing depression symptom measurement and for informing and advancing clinical decision-making, next-best actions, and personalized treatment recommendations. Moving analysis of speech for the detection of depression symptoms – not long ago deemed science fiction – to clinical reality presents considerable opportunities for changing the paradigm of BH care to be more efficient, personalized, proactive, and upstream-focused.

18

Acknowledgments

Author Contributions

Conceptualization: BEK, RJA, MA, ES, MH, Methodology: ES, BEK, MA, MH, RJA, Investigation: ES, BEK, MA, ES, MH, RJA, Supervision: BEK, MA, ES, MH, DH, RJA, Writing—original draft: BEK, MA, ES, MH, RJA, Writing—review & editing: BEK, DH, RJA, MA, ES. The authors would like to acknowledge Nina Roth for her extensive support in manuscript preparation.

Conflicts of Interest

MA, ES and MH are affiliated with Ellipsis Health, a for-profit healthcare technology company whose algorithms and analytic services were used in this study.

Data Availability

The datasets presented in this article are not readily available because many that support findings for this study are proprietary. Requests to access the datasets should be directed to MA, mike@ellipsishealth.com.

Abbreviations

ACM: Amazon Comprehend Medical
ASR: Automatic Speech Recognition
AWS: Amazon Web Services
BH: Behavioral Health
CCC: Concordance Correlation Coefficient
CM: Case Management
EER Equal Error Rate
IRB: Institutional Review Board
MAE: Mean Absolute Error
m: Mean
ML: Machine Learning
NLP: Natural Language Processing
PHQ-8: Patient Health Questionnaire, 8 questions
PHQ-9: Patient Health Questionnaire, 9 questions
ROC-AUC: Receiver Operating Curve-Area Under the Curve
sd: Standard Deviation
sens: Sensitivity

19

SFTP: Secure File Transfer Protocol
spec: Specificity
SVI: Social Vulnerability Index

Multimedia Appendix 1: Metrics explanation and model ROC performance

20

References

1. Substance Abuse and Mental Health Services Administration. 2021 NSDUH Annual National Report. CBHSQ Data. 2021. URL: https://www.samhsa.gov/data/report/2021-nsduh-annual-national-report [accessed 2024-06-24].

2. Vahratian A, Blumberg SJ, Terlizzi EP, Schiller JS. Symptoms of anxiety or depressive disorder and use of mental health care among adults during the COVID-19 pandemic — United States, August 2020–February 2021. MMWR Morb Mortal Wkly Rep. 2021;70:490-494. doi: 10.15585/mmwr.mm7013e2

3. Ettman CK, Abdalla SM, Cohen GH, Sampson L, Vivier PM, Galea S. Prevalence of Depression Symptoms in US Adults Before and During the COVID-19 Pandemic. JAMA Netw Open. 2020;3(9):e2019686. doi: 10.1001/jamanetworkopen.2020.19686

4. Wang PS, Demler O, Kessler RC. Adequacy of treatment for serious mental illness in the United States. Am J Public Health. 2002;92(1):92–98. doi: 10.2105/AJPH.92.1.92

5. Lovett L. Highmark health's behavioral health director: personalized care, upstream prevention will define the industry. BH Business. 2023 April 10. URL: https://bhbusiness.com/2023/04/10/highmark-healths-behavioral-health-director-personalized-care-upstream-prevention-will-define-the-industry/ [accessed 2024-06-24].

6. Fahed M, McManus K, Vahia IV, Offodile AC. Digital phenotyping of behavioral symptoms as the next frontier for personalized and proactive cancer care. JCO Clin Cancer Inform. 2022;6:e2200095. doi: 10.1200/CCI.22.00095

7. Huckvale K, Venkatesh S, Christensen H. Toward clinical digital phenotyping: a timely opportunity to consider purpose, quality, and safety. npj Digit Med. 2019;2:88. doi: 10.1038/s41746-019-0166-1

8. Insel TR. Digital phenotyping: technology for a new science of behavior. JAMA. 2017;318(13):1215–1216. doi: 10.1001/jama.2017.11295

9. Mohr DC, Shilton K, Hotopf M. Digital phenotyping, behavioral sensing, or personal sensing: names and transparency in the digital age. npj Digit Med. 2020;3:45. doi: 10.1038/s41746-020-0251-5

10. Chia AZR, Zhang MWB. Digital phenotyping in psychiatry: A scoping review. Technol Health Care. 2022;30:1331–1342. doi: 10.3233/THC-213648

11. Herrman H, Patel V, Kieling C, Berk M, Buchweitz C, Cuijpers P, Furukawa TA, Kessler RC, Kohrt BA, Maj M, McGorry P, Reynolds III CF, Weissman MM, Chibanda D, Dowrick C, Howard LM, Hoven CW, Knapp M, Mayberg HS, Penninx BWJH, Wolpert M. Time for united action on depression: a Lancet-World Psychiatric Association Commission. Lancet. 2022;399:957–1022. doi: 10.1016/S0140-6736(21)02141-3

12. Galatzer-Levy IR, Onnela JP. ML and the digital measurement of psychological health. Annu Rev Clin Psychol. 2023;19:133–154. doi: 10.1146/annurev-clinpsy-080921-073212

13. Low DM, Bentley KH, Ghosh SS. Automated assessment of psychiatric disorders

21

using speech: A systematic review. Laryngoscope Investig Oto. 2020;5:96–116. doi: 10.1002/lio2.354

14. Koutsouleris N, Hauser TU, Skvortsova V, De Choudhury M. From promise to practice: towards the realisation of AI-informed mental health care. Lancet Digit Health. 2022;4:e829–e840. doi: 10.1016/S2589-7500(22)00153-4

15. Flanagan O, Chan A, Roop P, Sundram F. Using acoustic speech patterns from smartphones to investigate mood disorders: scoping review. JMIR Mhealth Uhealth. 2021;9:e24352. doi: 10.2196/24352

16. Shen Y, Yang H, Lin L. Automatic depression detection: an emotional audio-textual corpus and a GRU/BiLSTM-based model. In: ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2022 May 22-27; Singapore, Singapore. IEEE; 2022. p. 6247-6251. doi: 10.1109/ICASSP43922.2022.9746569

17. Cohen AS, Cox CR, Le TP, Cowan T, Masucci MD, Strauss GP, Kirkpatrick B. Using machine learning of computerized vocal expression to measure blunted vocal affect and alogia. npj Schizophr. 2020;6:26. doi: 10.1038/s41537-020-00115-2

18. El-Den S, Chen TF, Gan YL, Wong E, O'Reilly CL. The psychometric properties of depression screening tools in primary healthcare settings: A systematic review. J Affect Disord. 2018;225:503–522. doi: 10.1016/j.jad.2017.08.060

19. Kroenke K, Spitzer RL, Williams JB. The PHQ-9: Validity of a brief depression severity measure. J Gen Intern Med. 2001;16:606–613. doi: 10.1046/j.1525-1497.2001.016009606.x

20. Kroenke K, Spitzer RL. The PHQ-9: A new depression diagnostic and severity measure. Psychiatr Ann. 2002;32:509–515. doi: 10.3928/0048-5713-20020901-06

21. Rutowski T, Shriberg E, Harati A, et al. Cross-demographic portability of deep NLP-based depression models. In: 2021 IEEE Spoken Language Technology Workshop (SLT); 2021 Dec 13-16; Shenzhen, China. IEEE; 2021. p. 1052–1057. doi: 10.1109/SLT48900.2021.9383609

22. Harati A, Shriberg E, Rutowski T, et al. Speech-based depression prediction using encoder-weight-only transfer learning and a large corpus. In: 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2021 Jun 6-11; Toronto, Canada. IEEE; 2021. p. 7273–7277. doi: 10.1109/ICASSP39728.2021.9414208

23. Lin D, Nazreen T, Rutowski T, et al. Feasibility of a ML-based smartphone application in detecting depression and anxiety in a generally senior population. Front Psychol. 2022;13:811517. doi: 10.3389/fpsyg.2022.811517

24. Baevski A, Zhou Y, Mohamed A, Auli M. wav2vec 2.0: a framework for self-supervised learning of speech representations. Advances in Neural Information Processing Systems. 2020;33:12449-1260. doi: 10.48550/arXiv.2006.11477

25. Beltagy I, Peters ME, Cohan A. Longformer: the long-document transformer. arXiv preprint arXiv:2004.05150. 2020 Apr 10. URL: https://arxiv.org/abs/2004.05150

26. Amazon Web Services, Inc. Speech to text - Amazon Transcribe - AWS. 2024 Jun 24. URL: https://aws.amazon.com/transcribe/ [accessed 2024-06-24].

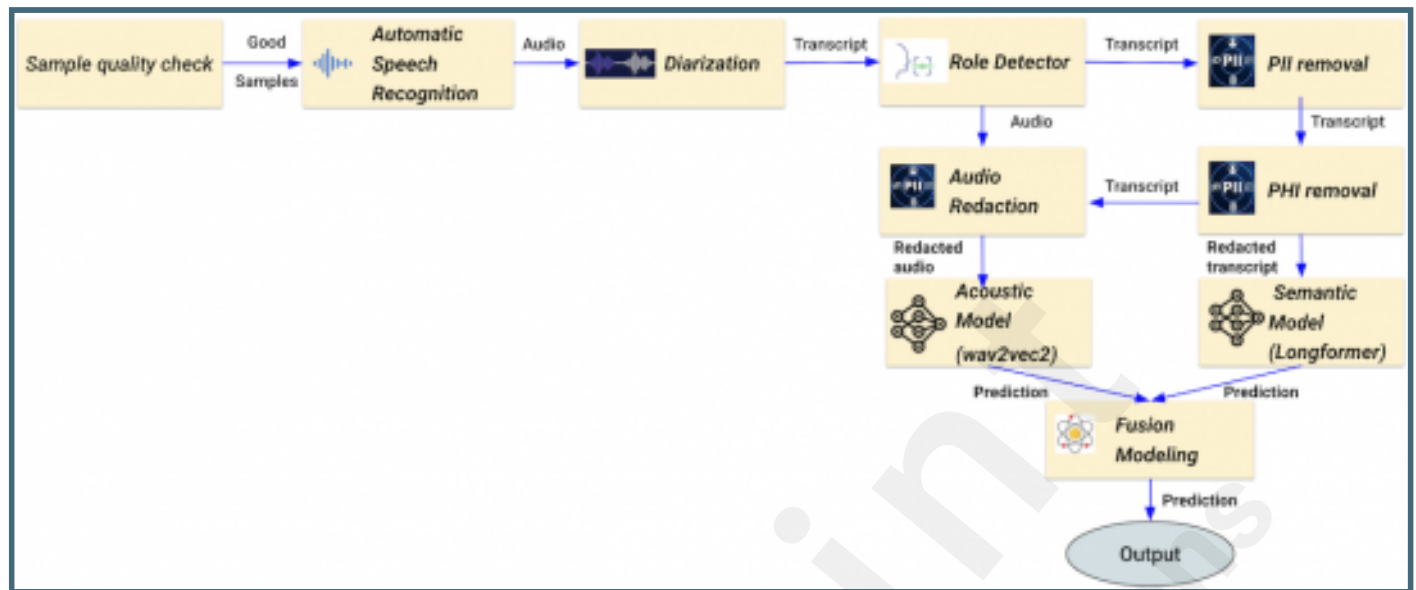27. Amazon Web Services, Inc. Introducing medical language processing with Amazon

22

Comprehend Medical | AWS ML blog. 2018 Nov 27. URL: https://aws.amazon.com/blogs/machine-learning/introducing-medical-language-processing-with-amazon-comprehend-medical/ [accessed 2024-06-24].

28. Fara S, Goria S, Molimpakis E, Cummins N. Speech and the n-Back task as a lens into depression. How combining both may allow us to isolate different core symptoms of depression. In: Interspeech 2022. ISCA; 2022. p. 1911-1915. doi: 10.48550/arXiv.2204.00088

29. CDC/ATSDR Social Vulnerability Index (CDC/ATSDR SVI). 2024 Jun 14. URL: https://www.atsdr.cdc.gov/placeandhealth/svi/index.html [accessed 2024-06-24].

30. Bn S, Abdullah S. Privacy-sensitive speech analysis using federated learning to assess depression. In: ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2022 May 23. IEEE; 2022. p. 6272-6276. doi: 10.1109/ICASSP43922.2022.9746827

31. Rutowski T, Harati A, Shriberg E, Lu Y, Chlebek P, Oliveira R. Toward corpus size requirements for training and evaluating depression risk models using spoken language. In: Interspeech 2022. ISCA; 2022. p. 3343-3347. doi: 10.21437/Interspeech.2022-10888

32. Seneviratne N, Espy-Wilson C. Multimodal depression severity score prediction using articulatory coordination features and hierarchical attention based text embeddings. In: Interspeech 2022. ISCA; 2022. p. 3353–3357. doi: 10.21437/Interspeech.2022-11099

33. Wang J, Ravi V, Flint J, Alwan A. Unsupervised instance discriminative learning for depression detection from speech signals. In: Interspeech 2022. ISCA; 2022. p. 2018–2022. doi: 10.21437/Interspeech.2022-10814

34. Zenebe Y, Akele B, W/Selassie M, Necho M. Prevalence and determinants of depression among old age: a systematic review and meta-analysis. Ann Gen Psychiatry. 2021;20:55. doi: 10.1186/s12991-021-00375-x

35. Devita M, De Salvo R, Ravelli A, De Rui M, Coin A, Sergi G, Mapelli D. Recognizing depression in the elderly: practical guidance and challenges for clinical management. Neuropsychiatr Dis Treat. 2022;18:2867-2880. doi: 10.2147/NDT.S347356

36. Gottfries CG. Is there a difference between elderly and younger patients with regard to the symptomatology and aetiology of depression? Int Clin Psychopharmacol. 1998;13:153-157. doi: 10.1097/00004850-199809005-00004

37. Hegeman JM, Kok RM, Van Der Mast RC, Giltay EJ. Phenomenology of depression in older compared with younger adults: Meta-analysis. Br J Psychiatry. 2012;200:275-281. doi: 10.1192/bjp.bp.111.095950

38. Ducat L, Philipson LH, Anderson BJ. The mental health comorbidities of diabetes. JAMA. 2014;312:691-692. doi: 10.1001/jama.2014.8040

39. Davenport S, Gray T, Melek S. Milliman High-Cost Patient Study 2020. Milliman Research Report. 2020 Aug 13. URL: https://www.milliman.com/-/media/milliman/pdfs/articles/milliman-high-cost-patient-study-2020.pdf [accessed 2024-06-24].

40. Bellon J, Quinlan C, Taylor B, Nemecek D, Borden E, Needs P. Association of outpatient behavioral health treatment with medical and pharmacy costs in the first

23

27 months following a new behavioral health diagnosis in the US. JAMA Netw Open. 2022;5:e2244644. doi: 10.1001/jamanetworkopen.2022.44644

41. Malpass A, Dowrick C, Gilbody S, Robinson J, Wiles N, Duffy L, Lewis G. Usefulness of PHQ-9 in primary care to determine meaningful symptoms of low mood: a qualitative study. Br J Gen Pract. 2016;66:e78–e84. doi: 10.3399/bjgp16X683473

42. Thombs BD, Kwakkenbos L, Levis AW, Benedetti A. Addressing overestimation of the prevalence of depression based on self-report screening questionnaires. CMAJ. 2018;190: E44–E49. doi: 10.1503/cmaj.170691

43. De Jong MG, Fox JP, Steenkamp JBE. Quantifying under- and overreporting in surveys through a dual-questioning-technique design. J Mark Res. 2015;52(6):737–753. URL: https://www.jstor.org/stable/43832399 [accessed 2024-06-24].

44. Hunt J, Eisenberg D. Mental health problems and help-seeking behavior among college students. J Adolesc Health. 2010;46(1):3–10. doi: 10.1016/j.jadohealth.2009.08.008

45. Ma S, Kang L, Guo X, et al. Discrepancies between self-rated depression and observed depression severity: The effects of personality and dysfunctional attitudes. Gen Hosp Psychiatry. 2021;70:25–30. doi: 10.1016/j.genhosppsych.2020.11.016

46. Harati A, Rutowski T, Lu Y, et al. Generalization of deep acoustic and NLP models for large-scale depression screening. In: Obeid I, Picone J, Selesnick I, editors. Biomedical Sensing and Analysis. Cham: Springer International Publishing; 2022. p. 99–132. doi: 10.1007/978-3-030-99383-2_318

47. Levis B, Benedetti A, Thombs BD; DEPRESsion Screening Data (DEPRESSD) Collaboration. Accuracy of Patient Health Questionnaire-9 (PHQ-9) for screening to detect major depression: individual participant data meta-analysis. BMJ. 2019 Apr 9;365. doi: 10.1136/bmj.l1476. Erratum in: BMJ. 2019 Apr 12;365. doi: 10.1136/bmj.l1781. PMID: 30967483; PMCID: PMC6454318

24

Supplementary Files

Figures

Deep learning architecture and processing pipeline.

Multimedia Appendixes

Metrics Explanation and Model ROC Performance.
URL: http://asset.jmir.pub/assets/95aae5012c042604ed1045819d520fd0.docx