EMU: Early Mental Health Uncovering Framework and Dataset

ML Tlachac, Ermal Toto, Joshua Lovering, Rimsha Kayastha, Nina Taurich, Elke Rundensteiner
Data Science and Computer Science Departments
Worcester Polytechnic Institute, Worcester, MA, USA
Emails: {mltlachac,toto,jlovering,rkayastha,ntaurich,rundenst}@wpi.edu

***Abstract*—Mental illnesses are often undiagnosed, demonstrating need for an effective unbiased alternative to traditional screening surveys. For this we propose our Early Mental Health Uncovering (EMU) framework that supports near instantaneous mental illness screening with non-intrusive active and passive modalities. We designed, deployed, and evaluated the EMU app to passively collect retrospective digital phenotype data and actively collect short voice recordings. Additionally, the EMU app also administered depression and anxiety screening surveys to produce depression and anxiety screening labels for the data. Notably, more than twice as many participants elected to share scripted audio recordings than any passive modality. We then study the effectiveness of machine learning models trained with the active modalities. Using scripted audio, EMU screens for depression with F1=0.746, anxiety with F1=0.667, and suicidal ideation with F1=0.706. Using unscripted audio, EMU screens for depression with F1=0.691, anxiety with F1=0.636, and suicidal ideation with F1=0.667. Jitter is an important feature for screening with scripted audio, while Mel-Frequency Cepstral Coefficient is an important feature for screening with unscripted audio. Further, the frequency of help-related words carried a strong signal for suicidal ideation screening with unscripted audio transcripts. This research results in a deeper understanding of the selection of modalities and corresponding features for mobile screening. The EMU dataset will be made available to public domain, representing valuable data resource for the community to further advance universal mental illness screening research.**

***Index Terms*—mobile health, depression screening, anxiety screening, audio classification, machine learning**

I. Introduction

Mental illnesses are prevalent, directly impacting around 20 percent of U.S. adults [1]. The two most prevalent mental illnesses are depression and anxiety which globally account for one trillion US dollars per year in lost productivity [1]. While depression is one of the most treatable mental illness [2], many people remain undiagnosed [3]. On average, it takes 11 years from mental illness symptom onset to receive treatment [1], which is partially due to delays in diagnosis. Lack of symptom recognition and medical resources are known to interfere with timely diagnosis [3]. Further, people with depression are less likely to seek help and more likely to delay seeking help [4].

Screening for mental illnesses is a national priority with the U.S. Preventative Services Task Force (USPSTF) endorsing screening all adults for depression [5]. Primarily, explicit surveys are used as instruments to screen for mental illnesses. Unfortunately, these surveys are often considered cumbersome and intrusive [6]. They also require honest self-reflection over the past two weeks. For these reason, they are highly susceptible to both conscious and unconscious bias. Thus, to ensure everyone is connected to needed resources, it is important to screen for mental illness in an unbiased manner.

In recent years, a number of alternative methods for screening have been proposed, including voice recordings [7], social media [8], [9], and smartphone sensor data [10]. While the Mood Assessment Capable Framework (Moodable) used multiple modalities [11], the focus was on passive modalities with the only active modality being a scripted voice recording. Yet, participants were most willing to share this voice recording.

Thus, to aid in the ultimate goal of universal screening, we thus propose our Early Mental Health Uncovering (EMU) framework for mental illness screening with both active and passive modalities. EMU represents a low burden and low bias approach towards almost instantaneous screening with whatever modalities participants were comfortable sharing. This allows us to study not only which modalities the participants are most willing to share, but also the relative predictive power of alternate modalities. EMU is unique in that it seamlessly combines the collection of both active and passive modalities to maximize the data collected for near instantaneous screening of a variety of mental illnesses.

To test the feasibility of the EMU screening framework, we develop and deploy the EMU mobile app to collect a variety of modalities including scripted audio, unscripted audio, text logs, call logs, GPS locations, and Twitter posts. The EMU app also administered mental illness screening surveys for depression and anxiety to label the data, producing a valuable data resource for machine learning. To validate our design choices, we report on our findings, including relative willingness of participants to provide each modality. We further use the most shared modalities to screen for depression, anxiety, and suicidal ideation. Our contributions include:

- A unique approach for near instantaneous collection of passive and active modalities for mental health screening.
- A publicly available dataset of digital phenotype and audio features with rich mental illness screening labels.
- A comparative study of the relative utility of scripted and unscripted mobile recordings for mental illness screening.

This work was supported by the US Department of Education P200A150306 & P200A180088: GAANN grants, NSF III: Small #1910880, and the Data Science department at WPI. Results were obtained in part using a HPC system acquired through NSF MRI grant DMS-1337943 to WPI.

- An exploration of important features for mental illness screening with scripted and unscripted audio.

II. Related Work on Mental Illness Screening

Screening using explicit screening surveys. While the USPSTF does not state a preferred depression screening tool, it mentions the Patient Health Questionnaire (PHQ) is among the most popular [5]. The full nine question version is referred to as the PHQ-9 [12]. When the ninth item regarding suicidal ideation is absent, the survey is known as the PHQ-8. The USPSTF further suggests anyone who screens positive for depression also be screened for comorbid conditions such as anxiety [5]. The seven question Generalized Anxiety Disorder-7 (GAD-7) [13] is the anxiety counterpart to the PHQ-9. Each question in these surveys measures symptom severity with a Likert scale ranging between 0 and 3. For these surveys, a collective score of at least 10 aggregated over all questions is clinically significant. At this cutoff, when compared to clinician expertise, the PHQ-9 as instrument has sensitivity and specificity of 88% [12], while the GAD-7 has sensitivity of 89% and specificity of 82% [13]. Unfortunately, not everyone is willing to complete these screening surveys as they are considered cumbersome and intrusive [6].

Screening using social media data. Surveys of studies focusing on depression detection with social media reveal that Twitter is the most popular social media platform for such research [8], [9]. These studies identified depression through multiple channels such as self declaration and screening surveys. Support Vector Machine (SVM) was the most common machine learning algorithm followed by logistic regression and random forests [9]. Other social media platforms are also occasionally used for depression detection research. For example, one study collected PHQ-9 scores from 165 new mothers to predict depression from Facebook posts [14]. Another study asked 749 crowd-sourced participants complete the PHQ-8 to predict depression with Instagram data [15].

Screening using audio recordings. A review of depressed and suicidal speech databases focuses on the impact of these conditions on common paralinguistic speech characteristics [7]. The most notable of the mentioned datasets is the publicly available Distress Analysis Interview Corpus Wizard-of-Oz (DAIC-WOZ) which contains 189 PHQ-8 labeled clinical interviews [16], [17]. During the Audio/Visual Emotion Challenge and Workshop (AVEC) in 2016 [18], the baseline results for depression classification using an SVM with DAIC-WOZ audio was $F1 = 0.41$. While such interviews yield large quantities of unscripted audio, they are time intensive and ask invasive questions. To address this limitation, other researchers collected short utterances and PHQ-9 scores with smartphones from 887 participants [19]. Due to the data being unfortunately unbalanced, their SVM screened for depression with a relatively low $F1 = 0.42$ score. While accuracy is somewhat higher at 0.73, this metric is known not to be meaningful in unbalanced datasets. Further, a recent study analyzed the correlation between sporadic smartphone recordings of environmental audio and mental illnesses assessed by the GAD-7 and PHQ-8 for 84 crowd-sourced participants [20].

Screening using smartphone sensor data. StudentLife was the first continuous sensing Android app to assess mental health and academic success [10]. This app was deployed to track 48 college students over 10 weeks. Correlations were calculated between features extracted from the sensing data and PHQ-9 scores [10]. Another study asked 28 participants to complete the PHQ-9 and carry a phone with an Android sensor data collection app for two weeks [21]; a correlation analysis was conducted. LifeRhythm, a cross-platform app, was deployed to collect location, activity data, and PHQ-9 scores from 79 college students over six months [22]. On this LifeRhythm dataset, an SVM model was shown to screen for depression based on PHQ-9 labels with $F1 = 0.52$.

Screening using multi-modal data. The Multi-modal Open Dataset for Mental-disorder Analysis (MODMA) contains 128-electrode and 3-electrode electroencephalograms (EEGs) from 55 patients as well as scripted and unscripted voice recordings with PHQ-9 labels from 52 patients at Lanzhou University Second Hospital in China [23]. An SVM trained on audio from all participants achieved average accuracies of 0.62 and 0.57 for scripted and unscripted audio, respectively [24]. In contrast, the Moodable study elected for a more passive approach by deploying an Android app to collect phone data from 335 MTurk participants for depression assessment [11]. Additionally, the participants were asked to record the scripted sentence "The quick brown fox jumps over the lazy dog". Based on PHQ-9 responses, random forest models screened for depression with $F1 = 0.58$ and accuracy $= 0.58$ as well as suicidal ideation with $F1 = 0.58$ and accuracy $= 0.62$ [11].

III. EMU Data Preparation and Machine Learning

The EMU approach leverages active and passive modalities to screen for mental illnesses in an efficient unbiased fashion. The EMU framework, in Fig. 1, has a client-server architecture with data collector, data server, model designer, and mental illness screener components. Specifically, the EMU approach is designed to support the collection of four types of data: surveys, retrospective smartphone data, social media data, and active prompts. As the smartphone data and social media posts are collected retrospectively, they are unable to be influenced by knowledge of this study and are therefore unbiased. Active

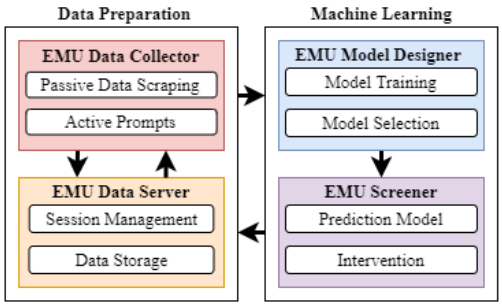

Fig. 1. The EMU framework to screen for mental illnesses.

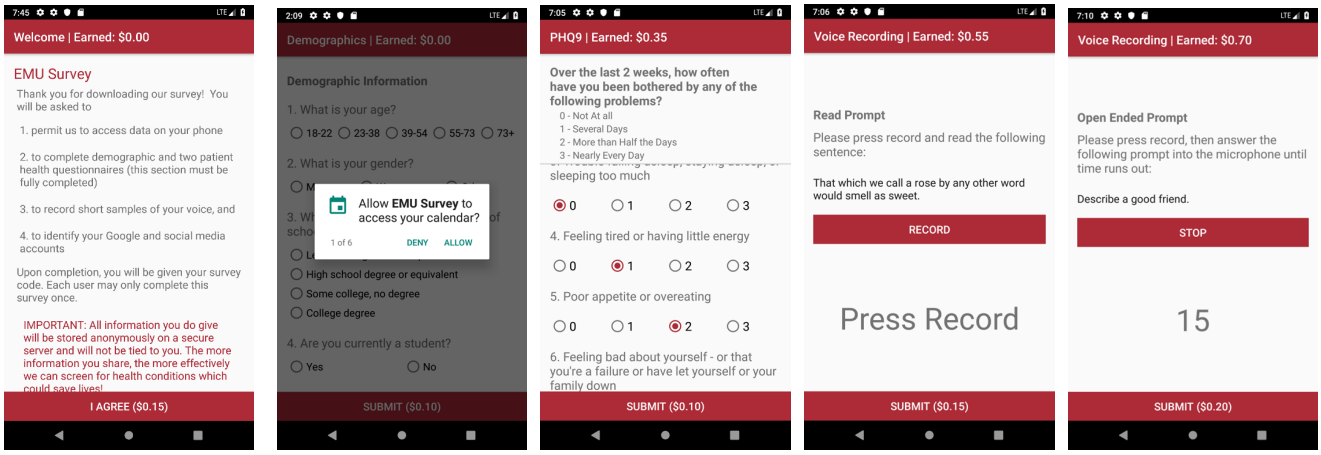

Fig. 2. Screenshots from the EMU data collection app.

prompts are intended to be quickly completed and designed to mitigate any potential bias created by screening awareness.

A. EMU Data Preparation: Collector and Server

For this feasibility study, we implemented the EMU data collector with a native Android app with multiple data collection units to encompass different types of data (Fig. 1). Namely, the Android OS Permission System is used to access retrospective smartphone data and external remote authentications systems access retrospective social media data. The EMU app is also designed to collect active prompts such as voice recordings. Lastly, the EMU app can administer surveys including traditional mental illness screening surveys which provides participant labels for tuning machine learning models.

Fig. 2 shows the participant-facing design of the EMU app by displaying interaction screenshots as the participant proceeds through the collection process. All sections of the app, in order, are displayed in Table I. The app first asks for permissions to collect the phone data so the scraping occurs as participants complete the remainder of the app.

Next, the EMU app administers the screening surveys as the data is useless for machine learning without labels. The screening surveys are the PHQ-9 [12] for depression and GAD-7 for anxiety [13]. The PHQ-9 was used by many studies in Sec. II and has been shown to be effective regardless of mode of administration [25]. The GAD-7 is the anxiety counterpart to the PHQ-9 with both screening surveys ask participants to reflect on the prior two weeks when responding to the questions regarding symptom severity. To prevent survey fatigue, we only ask four multiple-choice demographic questions regarding age, gender, education, and student status.

Given our hypothesis that most people will share active prompts, we place them next in the collection process. We select audio as the mode for the active prompts as voice is a rich modality that requires little time or effort from participants. Further, screening awareness is unlikely to impact the predictive quality of voice as it is difficult to alter vocal features to match a screening outcome. The EMU app collects both scripted and unscripted audio prompts from participants. We select audio prompts that we hypothesize will elicit emotional responses and thus be more useful when screening for mental illnesses. For the scripted audio, participants are prompted to repeat the iconic Shakespeare quote "that by which we call a rose by any other word would smell as sweet". For the unscripted audio, participants are prompted to "describe a good friend". This open-ended prompt is intentionally vague to elicit varied interpretations as we hypothesize varied transcript content will be more valuable for screening purposes.

The last part of the EMU collection process contains the external remote authentication systems since participants may not have these external accounts nor know their login information. While sharing this data was optional, we did not want to discourage participants earlier in the collection process. Specifically, participants are asked to share their Twitter username and login to their Google Maps and Instagram accounts.

All data collected by the EMU app is TSL encrypted before being sent to a secure server for storage. The EMU data sever is also responsible for session management. It stores the smartphone's hardware identifier (phoneID) to easily distinguish unique participants. For this deployment, when a session is initiated, the phoneID is compared to those in completed sessions, thus allowing for a timely repetition notification.

B. Crowd-Sourced Data Collection Study Using EMU

To collect a rich dataset of labeled voice recordings and digital phenotype data, we deployed the aforementioned EMU mobile collection app. Participants were recruited through Amazon's Mechanical Turk (MTurk) crowdsourcing platform [26]. The compensation in Table I reflects both the time and effort required to complete each section. Only the demographic survey and PHQ-9 were mandatory. A participant who completed all sections, estimated to take at most six minutes, would earn $1.10 US dollars ($11 per hour). If participants attempted to repeat the collection with the same phone, they were informed "You have already completed a full session". Data was collected from 2/18/2019 to 3/13/2019 under [IRB number redacted for double-blind submission].

TABLE I
ORDER DATA WAS COLLECTED BY THE EMU APP WITH THE ESTIMATED COMPLETION TIME AND THE REWARD FOR EACH SECTION.

Section	Time in minutes	Reward in US dollars
Introduction	0.25	\$0.15
Phone Permissions	0.25	\$0.10
Demographics	0.5	\$0.10
PHQ-9	1	\$0.10
GAD-7	1	\$0.10
Scripted Audio	0.5	\$0.15
Unscripted Audio	0.75	\$0.20
Google Maps	0.5	\$0.10
Twitter	0.25	\$0.05
Instagram	0.5	\$0.05
Total	6	\$1.10

C. EMU Machine Learning: Model Designer and Screener

Our aim is to compare the relative utility of the most shared modalities when screening for participant depression, anxiety, and suicidal ideation. As hypothesized, the active modalities were indeed the most shared modalities (Table II). In fact, scripted and unscripted audio were the only modalities shared by more than half of the participants and therefore are most useful for machine learning. Below, we now compare their mental illnesses screening ability by training machine learning models to predict depression (PHQ-9$\geq$ 10) [2], anxiety (GAD-7$\geq$ 10) [13], and suicidal ideation (PHQ-9 item-9$\geq$ 1) [2].

Feature engineering in EMU. We apply feature engineering to extract 1582 features from each audio recording using openSMILE [27], the same toolkit used to extract audio features for AVEC 2016 [18]. Unlike scripted audio, the transcripts of the unscripted audio vary in content. Thus, from these transcripts, we further extract text features by replicating the text feature engineering from a prior study that successfully screened for depression with self-written text [28]. The 231 text features include word category frequencies with Empath [29], part of speech tag frequencies and sentiment with TextBlob [30], and volume (number of words and characters).

Feature selection in EMU. We experiment with three common feature selection and/or reduction techniques: principal component analysis (PCA), chi-squared, and Extra-Trees. PCA calculates successive linear combinations, known as principal components, from the original features such that each principal component explains the maximum amount of variance not already explained [31]. The chi-squared statistic measures the dependence between each feature and the target variable [31]. Chi-squared selected features are those with the highest chi-squared statistics. The Extra-Trees algorithm, a variation of the random forest algorithm [31] [32], is a popular tree-based algorithm for selecting features for audio classification [33].

Machine learning methods in EMU. Traditional machine learning methods are preferred in healthcare [34] due to their interpretability and applicability to smaller datasets. Thus, we compare a representative selection of machine learning models: support vector classifier (SVC), Gaussian Naive Bayes classifier (NB), logistic regression (LR), k-nearest neighbor (kNN), and random forest (RF) [31]. We also consider Extreme Gradient Boosting (XGBoost) [35], a more advanced tree-based algorithm. We leverage both linear and Gaussian kernels for SVC models. After a preliminary exploration, we noted other parameters had little effect and used the defaults.

Metrics for EMU screening model evaluation. We employ a leave-one-out strategy for evaluation which is ideal for robust evaluation on smaller datasets. This is a form of cross-validation where the training data consists of the features for all but one participant. The feature selection/reduction technique is applied to the training data with the features normalized between 0 and 1. The training data is then upsampled to balance classes. As each test set consists of the features for only one participant, each model makes a single prediction for that participant. We then evaluate each model configuration by considering the number of true positive (tp), false positive (fp), false negative (fn), and true negative (tn) predictions.

Given that $F1 = (2tp/(2tp + fp + fn))$ focuses on *true positive predictions*, we consider the best model configurations to be those that maximize $F1$. We also report *sensitivity* ($tp/(tp + fn)$) and *specificity* ($tn/(tn + fp)$) for their diagnostic usefulness. Further, we report AUC and accuracy for comparison purposes given their popularity in related studies [8], [23]. AUC is the aggregated performance across different thresholds on the ROC curve which is formed by plotting sensitivity against the false positive rate ($fp/(fp+tn)$). Accuracy simply denotes the ratio of correctly classified participants and is therefore considered less useful in health domains.

IV. AVAILABILITY

The featurized EMU dataset and the machine learning code are available at *https://github.com/mltlachac/EMU*. We replicated the aforementioned text feature engineering on the text messages and tweets to obtain features for these modalities.

V. RESULTS

During the three weeks the EMU data collection app was posted on MTurk, 70 unique participants completed the mandatory demographics and PHQ-9; 60 of them completed the entire data collection. The number of completed, repeated, and incompleted sessions are displayed in Fig. 3. Based on the demographic survey, the 70 participants consisted of 54 men and 16 women. 14 reported being students. Most reported to be in the 23-38 age range and have a college degree. The distribution of PHQ-9 and GAD-7 scores are in Fig. 4.

A. Willingness to Share

As displayed in Table II, scripted audio was the most completed non-survey modality with 90% of participants submitting a recording. The only other predictive modality that more than 50% of the participants shared was unscripted audio. This confirms our hypothesis that people are more willing to share quick active modalities than their more private passive modalities. Phone modalities may be perceived as even more invasive than the traditional screening surveys and thus are not likely to result universal screening by themselves. This

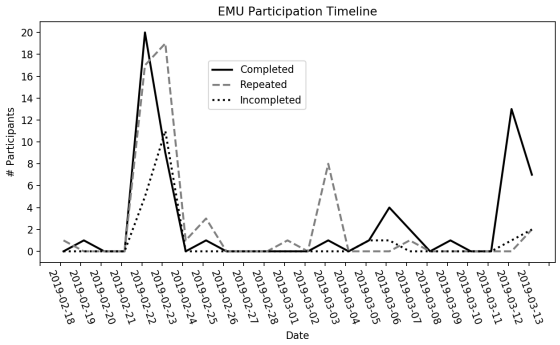

Fig. 3. Number of sessions of EMU data collection app.

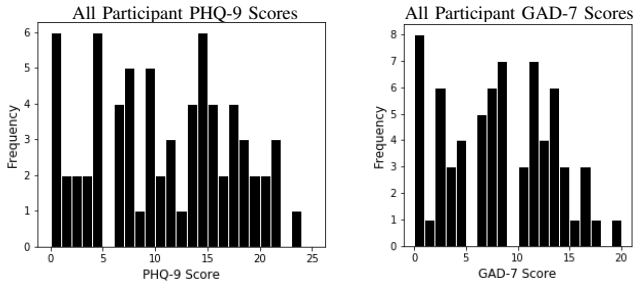

Fig. 4. Participant PHQ-9 and GAD-7 scores distributions.

TABLE II
NUMBER OF PARTICIPANTS WHO SHARED EACH MODALITY.

Modality	Participants	mean PHQ-9	mean GAD-7
PHQ-9	70	10.37	7.77
GAD-7	69	10.43	7.77
Demographic	70	10.37	7.77
Scripted Audio	63	10.08	7.63
Unscripted Audio	55	10.04	7.71
Text Messages	31	10.29	8.20
Calendar	28	10.89	8.18
Call Log	25	10.88	8.04
GPS	14	9.29	7.36
Twitter Username	14	8.50	7.21
Twitter Posts	11	6.91	6.00
Instagram	0	0	0

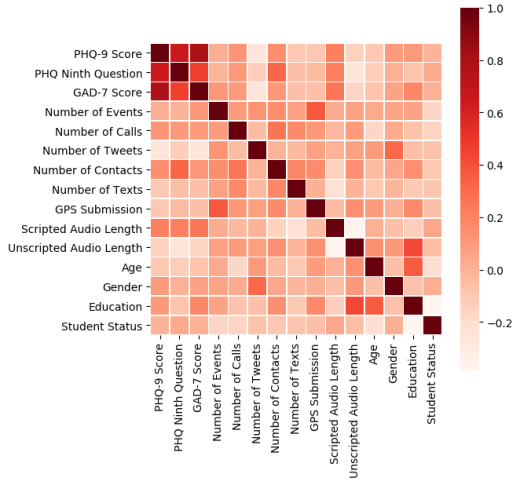

Fig. 5. Pearson correlations between submitted data.

validates our design choice to allow participants to submit both active and passive for mental illness screening based on their preferences. Surprisingly, social media was least likely to be shared, though this may reflect the number of participants with accounts rather than the willingness to share.

As asked, 56 of the 63 participants who submitted scripted audio recordings attempted to repeat the prompt and can be considered compliant. Participants who submitted unscripted audio interpreted the prompt different and thus had varied responses: 31 described qualities of a good friend, 11 described particular friends, 6 repeated the prompt verbatim, and 1 apologized for not having any close friends. The verbatim responses were likely due to confusion from the scripted audio preceding the unscripted audio. The remaining 6 played unrelated noise and are therefore considered uncompliant.

Table II also contains the mean PHQ-9 and GAD-7 scores for the subset of participants who submitted each modality. For all participants, the mean PHQ-9 score was 10.37 and the mean GAD-7 score was 7.77. According to t-tests, none of the participants subsets had statistically significantly different scores than all participants. While the subset who shared their Twitter usernames had lower scores than the subset who shared any other modalities, participants with tweets posted to their accounts had even lower mean PHQ-9 and GAD-7 scores.

B. Correlation Between Submitted Data and Mental Illness

Fig. 5 displays Pearson correlations between the mental illness screening scores and data features. The PHQ-9 and GAD-7 scores have the highest pairwise correlation coefficient of 0.78 which is very similar to the 0.75 correlation in the literature [13], validating the quality of the crowd-sourced EMU dataset. We extract the length from the audio recordings and quantity of passive modalities entries (except GPS) during the last two weeks. Missing values are replaced by 0 or the mean. We treat age and education as ordinal features. We also construct categorical features by considering who shared GPS, reported being students, and identified as women. The number of contacts and the ninth question regarding suicidal ideation have a correlation of 0.33. No other feature has a higher correlation with any of the mental illnesses.

C. Machine Learning Results for Mental Illness Prediction

As more than half of the participants shared scripted and unscripted audio, we train machine learning models to screen for depression (PHQ-9$\geq$ 10), anxiety (GAD-7$\geq$ 10), and suicidal ideation (item-9$\geq$ 1) with features extracted from the audio recordings. Given the gender disparity of depression [36] and existence of gender labels in comparative datasets [16], [23], we include self-reported gender in the set of features.

1) Scripted vs Unscripted Audio: The performance of the model configurations with the highest $F1$ for each mental illness and audio type are displayed in Table III. Models built with scripted audio had higher $F1$ scores than those built with unscripted audio. Depression proved to be the easiest mental illness to predict for both types of audio. Chi-squared feature selection proved best for selecting features from scripted

TABLE III

RESULTS OF THE MODEL CONFIGURATIONS WITH THE HIGHEST F1 SCORES USING AUDIO FEATURES. WHILE THE LEAVE-ONE-OUT EVALUATION STRATEGY DOES NOT YIELD STANDARD DEVIATIONS, IS IDEAL FOR SMALL DATASETS AS IT MAXIMIZES THE QUANTITY OF TRAINING DATA.

Mental Illness	Audio	Model	Selection	Features	$F1$	AUC	Acc	$Sens$	$Spec$	tp	fp	fn	tn
Depression	Scripted	Gaussian SVC	Chi-squared	3	0.746	0.729	0.730	0.781	0.677	25	10	7	21
Depression	Unscripted	Gaussian SVC	Extra-Trees	8	0.691	0.691	0.691	0.704	0.676	19	9	8	19
Anxiety	Scripted	Gaussian SVC	Chi-squared	2	0.655	0.696	0.683	0.760	0.632	19	14	6	24
Anxiety	Unscripted	NB	Extra-Trees	6	0.625	0.681	0.673	0.714	0.647	15	12	6	22
Suicidal Ideation	Scripted	Linear SVC	Chi-squared	10	0.706	0.752	0.762	0.692	0.811	18	7	8	30
Suicidal Ideation	Unscripted	kNN	Chi-squared	3	0.627	0.667	0.655	0.727	0.606	16	13	20	6

TABLE IV

RESULTS OF THE MODEL CONFIGURATIONS WITH THE HIGHEST $F1$ SCORES BUILT WITH AUDIO FROM THE COMPLIANT SUBSET OF PARTICIPANTS.

Mental Illness	Audio	Model	Selection	Features	$F1$	AUC	Acc	$Sens$	$Spec$	tp	fp	fn	tn
Depression	Scripted	Gaussian SVC	Chi-squared	4	0.746	0.735	0.732	0.814	0.655	22	10	5	19
Depression	Unscripted	kNN	PCA	2	0.667	0.675	0.673	0.696	0.654	16	9	7	17
Anxiety	Scripted	XGBoost	Chi-squared	8	0.667	0.733	0.750	0.667	0.800	14	7	7	28
Anxiety	Unscripted	Linear SVC	Chi-squared	8	0.636	0.695	0.673	0.778	0.613	14	12	4	19
Suicidal Ideation	Scripted	kNN	Chi-sqaured	10	0.648	0.729	0.768	0.571	0.886	12	4	9	31
Suicidal Ideation	Unscripted	Linear SVC	Extra-Trees	2	0.629	0.714	0.735	0.647	0.781	11	6	7	25

audio; only 3 and 2 chi-squared selected features were required to screen for depression and anxiety, respectively. For these mental illnesses, the models built with unscripted audio required more features to achieve the displayed $F1$ scores. Screening for suicidal ideation with scripted audio required the most features, indicating many different vocal features contain information useful for this classification task.

2) Compliant vs All Audio Participants: This experiment is to determine the impact of non-compliant participants, i.e., participants who submitted audio but did not try to adhere to the prompt. There are 56 scripted and 49 unscripted compliant recordings. The models configurations with the highest $F1$ scores for these recordings are displayed in Table IV. When screening for depression and suicidal ideation, the $F1$ score remains the same or decreases when compared to models in Table III. However, the unscripted audio models in Table IV only require 2 features. As the majority of non-compliant unscripted audio participants screened positive for depression, non-compliant responses may be indicative of certain mental illnesses, though further study must be conducted to ensure this finding is generalizable. In contrast, for anxiety screening, the $F1$ scores increased when using only compliant recordings but 8 features were required for both types of audio.

3) Audio vs Transcript Features: We extracted text and audio features from the 43 unscripted audio recordings that contained varying transcript content. This does not include the 6 uncompliant participants or the 6 participants who repeated the prompt verbatim. Table V compares the model configurations with the highest $F1$ scores. When screening for depression and anxiety, this subset of unscripted recordings yielded lower $F1$ scores when leveraging just audio features than the subsets of recordings in Tables III and IV. This suggests that the verbatim responses still contained valuable audio for screening purposes, though the lower screening scores may be due to the the experiments in Table V containing fewer participants or having a smaller ratio of positive instances. In contrast, this subset of unscripted audio recordings yielded higher $F1$ scores when screening for suicidal ideation, and only one principal component was required. Only for anxiety screening are the transcript features more predictive than the audio features. In no cases does combining the text and audio features increase the $F1$ scores for this dataset.

D. Important Audio Features for Machine Learning

The visual separability of participants with depression and anxiety based on the two most important selected scripted and unscripted audio features are displayed in Fig. 6. The features were selected from the audio features of all participants before any were removed to create training sets so not all models may have used these specific features. Local jitter is among the top two chi-squared selected scripted audio features for both depression and anxiety. Jitter, a feature known to be indicative of depression [37], is defined as the "cycle-to-cycle variation of fundamental frequency" and accounts for a voice sounding breathy or rough [38]. For the unscripted audio, Mel-Frequency Cepstral Coefficients (MFCCs) were among the top two features selected by the Extra-Trees algorithm to screen for both anxiety and depression. The MFCCs are extracted from the Mel frequency scale and are less impacted by natural speech variations than a speech waveform [39]. As Mel frequency scales are used to capture phonetically important characteristics of speech [39], it is logical that these features are valuable for screening with unscripted voice.

VI. DISCUSSION

A. Comparative Screening Capabilities with EMU Audio

When screening for depression, we achieved $F1$ scores of 0.746 with scripted audio features and 0.691 with unscripted audio features. These results are significantly higher than the $F1 = 0.41$ reported for the unscripted audio in AVEC 2016 [18] and the $F1 = 0.58$ reported for the scripted audio in Moodable in 2020 [11] despite also using openSMILE features. Deep learning models have since been used to improve

TABLE V

RESULTS OF THE MODEL CONFIGURATIONS WITH THE HIGHEST $F1$ SCORES BUILT WITH AUDIO AND/OR TRANSCRIPT FEATURES FROM UNSCRIPTED AUDIO RECORDINGS FROM THE SUBSET OF PARTICIPANTS WHO SHARED UNSCRIPTED AUDIO RECORDINGS WITH VARYING TRANSCRIPT CONTENT.

Mental Illness	Data	Model	Selection	Features	$F1$	AUC	Acc	$Sens$	$Spec$	tp	fp	fn	tn
Depression	Audio	Gaussian SVC	PCA	3	0.600	0.633	0.628	0.667	0.600	12	10	6	15
Depression	Transcript	Gaussian SVC	PCA	1	0.538	0.489	0.442	0.778	0.200	14	20	4	5
Depression	Both	Gaussian SVC	PCA	3	0.600	0.633	0.628	0.667	0.600	12	10	6	15
Anxiety	Audio	XGBoost	Extra-Trees	9	0.600	0.693	0.721	0.600	0.786	9	6	6	22
Anxiety	Transcript	kNN	PCA	1	0.629	0.706	0.698	0.733	0.679	11	9	4	19
Anxiety	Both	Gaussian SVC	Extra-Trees	5	0.545	0.639	0.651	0.600	0.679	9	9	6	19
Suicidal Ideation	Audio	Gaussian SVC	PCA	1	0.667	0.754	0.767	0.714	0.793	10	6	4	23
Suicidal Ideation	Transcript	Linear SVC	PCA	1	0.545	0.649	0.651	0.643	0.655	9	10	5	19
Suicidal Ideation	Both	Gaussian SVC	PCA	4	0.621	0.718	0.744	0.643	0.793	9	6	5	23

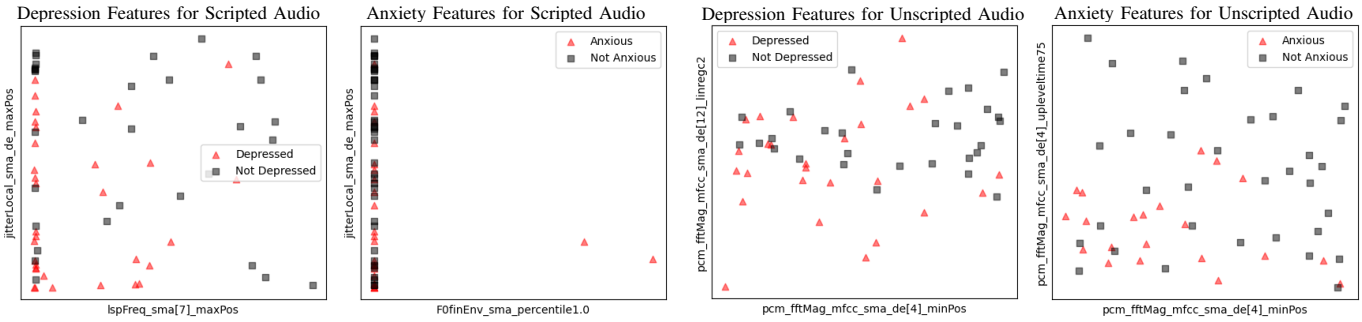

Fig. 6. Visual separability of participants based on the top two audio features selected for screening by the chi-squared statistic for scripted audio and the Extra-Trees algorithm for unscripted audio. Jitter and Mel-frequency Cepstral Coefficient (MFCC) are important for scripted and unscripted audio, respectively.

depression screening of DAIC-WOZ data to an $F1 = 0.63$ [40], which still is less than our current depression screening capabilities with EMU audio. Thus, our results may be a lower bound for EMU audio screening capabilities, though psychopathology still prefers traditional machine learning models due to their interpretability [41]. Like other research that collected voice recordings with smartphones [11], [19], the EMU audio was collected in natural environmental conditions. Thus, our higher screening capabilities are likely due to our audio prompts and/or feature selection strategies.

B. Crowd-sourced Participants have High Rates of Depression

The distribution of PHQ-9 and GAD-7 scores in Fig. 4 are notably not as left skewed as would be expected from the general population. A prior study [20] also noticed higher than normal rates of depression from participants on a crowd-sourced platform who completed their app-based data collection. Those researchers hypothesized that crowd-sourced workers experience higher than normal rates of mental illness and/or crowd-sourced workers with mental illness were more likely to engage with a mental illness study [20]. We propose a third hypothesis: crowd-sourced workers with mental illness are more apt to download a research app. Regardless, research indicates that crowd-sourcing platforms yield quality data [26].

C. The Future of Mental Illness Screening with EMU

The intent of the EMU framework is to replace explicit mental illness screening surveys with unobtrusively collected data. Of course, trained clinicians would ultimately still be required to confirm the diagnoses. Screening surveys have a number of limitations. For instance, the symptoms of depression may prevent people from seeking help [4] and disclosing depressive symptoms even if they want help [3]. The EMU framework could make the goal of effective universal screening a reality by offering a customizable, low burden, and unbiased solution.

The audio survey authors [7] stress the need for better collaboration in this domain. Their first recommendation is to share datasets, code, and transparent results. They also note that a more standardized approach to data collection is required [7]. EMU provides these needs to further screening research. While we only trained models on the most shared modalities, the less shared modalities could be combined with other similarly collected data to increase data quantity and improve study generalizability. For example, EMU and Moodable text logs were combined to screen for depression [42], [43].

VII. CONCLUSION

EMU is a framework for mental illness screening that integrates both passive and active modalities. We deployed our EMU app to collect voice recordings and digital phenotype data. The result is a valuable community data resource composed of digital phenotype and audio features labeled with mental illness scores. As participants were more willing to consistently share active than passive modalities, we trained machine learning models on features from the audio recordings to screen for mental illnesses. When using scripted audio features, the models screened for depression with $F1 = 0.746$, anxiety with $F1 = 0.667$, and suicidal ideation with $F1 = 0.706$. While using unscripted audio features, the models screened for depression with $F1 = 0.691$, anxiety with $F1 = 0.636$, and suicidal ideation with $F1 = 0.667$.

Future work involves exploring the mental illness screening capabilities of recent multimodal deep transfer learning models [44] on data from multiple existing datasets in this domain.

ACKNOWLEDGMENT

We thank Gao, Flannery, Resom, Assan, and Wu for helping collect the EMU dataset. We thank Gerych, Thadajarassiri, Buquicchio, Yin, and the rest of the DSRG at WPI for advice.

REFERENCES

[1] National Alliance on Mental Illness, "Mental health by the numbers," 2020. [Online]. Available: https://www.nami.org/mhstats

[2] K. Kroenke, R. L. Spitzer, and J. B. Williams, "The phq-9: validity of a brief depression severity measure," *Journal of General Internal Medicine*, vol. 16, no. 9, pp. 606–613, 2001.

[3] R. M. Epstein, P. R. Duberstein, M. D. Feldman *et al.*, "I didn't know what was wrong: How people with undiagnosed depression recognize, name and explain their distress," *Journal of General Internal Medicine*, vol. 25, pp. 954–961, 2010.

[4] K. Demyttenaere, A. Bonnewyn, R. Bruffaerts *et al.*, "Comorbid painful physical symptoms and depression: prevalence, work loss, and help seeking," *Journal of affective disorders*, vol. 92, pp. 185–193, 2006.

[5] A. Siu, K. Bibbins-Domingo, D. Grossman *et al.*, "Screening for depression in adults: Us preventive services task force recommendation statement," *Jama*, vol. 315, no. 4, pp. 380–387, 2016.

[6] M. D. Weist, M. Rubin, E. Moore, S. Adelsheim, and G. Wrobel, "Mental health screening in schools," *Journal of School Health*, vol. 77, no. 2, pp. 53–58, 2007.

[7] N. Cummins, S. Scherer, J. Krajewski *et al.*, "A review of depression and suicide risk assessment using speech analysis," *Speech Communication*, vol. 71, pp. 10–49, 2015.

[8] S. Guntuku, D. Yaden, M. Kern, L. Ungar, and J. Eichstaedt, "Detecting depression and mental illness on social media: An integrative review," *Current Opinion in Behavioral Sciences*, vol. 18, 2017.

[9] S. Chancellor and M. De Choudhury, "Methods in predictive techniques for mental health status on social media: a critical review," *NPJ digital medicine*, vol. 3, no. 1, pp. 1–11, 2020.

[10] R. Wang, F. Chen, Z. Chen *et al.*, "Studentlife: assessing mental health, academic performance and behavioral trends of college students using smartphones," in *the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 2014, pp. 3–14.

[11] A. Dogrucu, A. Perucic, A. Isaro *et al.*, "Moodable: On feasibility of instantaneous depression assessment using machine learning on voice samples with retrospectively harvested smartphone and social media data," *Smart Health*, pp. 100–118, 2020.

[12] K. Kroenke, R. Spitzer, and J. Williams, "The phq-9: validity of a brief depression severity measure," *Journal of General Internal Medicine*, vol. 16, no. 9, 2001.

[13] R. L. Spitzer, K. Kroenke, J. B. Williams, and B. Löwe, "A brief measure for assessing generalized anxiety disorder: the gad-7," *Archives of Internal Medicine*, vol. 166, no. 10, pp. 1092–1097, 2006.

[14] M. De Choudhury, S. Counts, E. J. Horvitz, and A. Hoff, "Characterizing and predicting postpartum depression from shared facebook data," in *the 17th ACM Conference on Computer Supported Cooperative Work Social Computing*, 2014, p. 626–638.

[15] B. J. Ricard, L. A. Marsch, B. Crosier, and S. Hassanpour, "Exploring the utility of community-generated social media content for detecting depression: An analytical study on instagram," *JMIR*, 2018.

[16] J. Gratch, R. Artstein, G. M. Lucas *et al.*, "The distress analysis interview corpus of human and computer interviews." in *Language Resources and Evaluation*. CiteSeer, 2014, pp. 3123–3128.

[17] D. DeVault, R. Artstein, G. Benn *et al.*, "Simsensei kiosk: A virtual human interviewer for healthcare decision support," in *the 2014 International Conference on Autonomous Agents and Multi-Agent Systems*, 2014, pp. 1061–1068.

[18] M. Valstar, J. Gratch, B. Schuller *et al.*, "Avec 2016: Depression, mood, and emotion recognition workshop and challenge," in *6th International Workshop on Audio/Visual Emotion Challenge*. ACM, 2016.

[19] Z. Huang, J. Epps, D. Joachim, and M. Chen, "Depression detection from short utterances via diverse smartphones in natural environmental conditions." in *INTERSPEECH*, 2018, pp. 3393–3397.

[20] D. Di Matteo, K. Fotinos, S. Lokuge *et al.*, "The relationship between smartphone-recorded environmental audio and symptomatology of anxiety and depression: Exploratory study," *JMIR Formative Research*, vol. 4, no. 8, 2020.

[21] S. Saeb, M. Zhang, C. J. Karr *et al.*, "Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: an exploratory study," *JMIR*, vol. 17, no. 7, 2015.

[22] A. Farhan, C. Yue, R. Morillo *et al.*, "Behavior vs. introspection: refining prediction of clinical depression via smartphone sensing data," in *IEEE Wireless Health*, 2016, pp. 1–8.

[23] H. Cai, Y. Gao, S. Sun *et al.*, "Modma dataset: a multi-model open dataset for mental-disorder analysis," *arXiv*, pp. arXiv–2002, 2020.

[24] Z. Liu, D. Wang, L. Zhang, and B. Hu, "A novel decision tree for depression recognition in speech," *arXiv preprint:2002.12759*, 2020.

[25] N. BinDhim, A. Shaman, L. Trevena *et al.*, "Depression screening via a smartphone app: cross-country user characteristics and feasibility," *Journal of the American Medical Informatics Association*, vol. 22, no. 1, pp. 29–34, 2014.

[26] M. Buhrmester, T. Kwang, and S. D. Gosling, "Amazon's mechanical turk: A new source of inexpensive, yet high-quality data?" *American Psychological Association*, 2016.

[27] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: the munich versatile and fast open-source audio feature extractor," in *the 18th ACM International Conference on Multimedia*, 2010, pp. 1459–1462.

[28] M. L. Tlachac and E. Rundensteiner, "Screening for depression with retrospectively harvested private versus public text," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 11, 2020.

[29] E. Fast, B. Chen, and M. S. Bernstein, "Empath: Understanding topic signals in large-scale text," in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 2016, pp. 4647–4657.

[30] S. Loria, "Textblob: Simplified text processing," 2018. [Online]. Available: https://textblob.readthedocs.io/en/dev/

[31] F. Pedregosa, G. Varoquaux, A. Gramfort *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[32] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Machine Learning*, vol. 63, no. 1, pp. 3–42, 2006.

[33] A. Jović, K. Brkić, and N. Bogunović, "A review of feature selection methods with applications," in *38th International Convention on Information and Communication Technology, Electronics and Microelectronics*. IEEE, 2015, pp. 1200–1205.

[34] D. B. Dwyer, P. Falkai, and N. Koutsouleris, "Machine learning approaches for clinical psychology and psychiatry," *Annual Review of Clinical Psychology*, vol. 14, pp. 91–118, 2018.

[35] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd ACD Sigkdd International conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794.

[36] K. H. Abate, "Gender disparity in prevalence of depression among patient population: a systematic review," *Ethiopian Journal of Health Sciences*, vol. 23, no. 3, pp. 283–288, 2013.

[37] S. Alghowinem, R. Goecke, M. Wagner *et al.*, "Detecting depression: a comparison between spontaneous and read speech," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2013, pp. 7547–51.

[38] M. Farrús, J. Hernando, and P. Ejarque, "Jitter and shimmer measurements for speaker recognition," in *Eighth annual conference of the international speech communication association*, 2007.

[39] R. Hasan, M. Jamil, G. Rahman, and S. Rahman, "Speaker identification using mel frequency cepstral coefficients," *variations*, vol. 1, no. 4, 2004.

[40] A. Mallol-Ragolta, Z. Zhao, L. Stappen, N. Cummins, and B. Schuller, "A hierarchical attention network-based approach for depression detection from transcribed clinical interviews," *Interspeech*, 2019.

[41] D. B. Dwyer, P. Falkai, and N. Koutsouleris, "Machine learning approaches for clinical psychology and psychiatry," *Annual review of clinical psychology*, vol. 14, pp. 91–118, 2018.

[42] M. L. Tlachac and E. Rundensteiner, "Depression screening from text message reply latency," in *42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 2020.

[43] M. L. Tlachac, V. Melican, M. Reisch, and E. Rundensteiner, "Mobile depression screening with time series of text logs and call logs," in *IEEE EMBS Intern. Conf. on Biomedical & Health Informatics (BHI)*, 2021.

[44] E. Toto, M. Tlachac, and E. Rundensteiner, "Audibert: A deep transfer learning multimodal screening framework for depression classification," in *30th ACM CIKM Applied Research Track*, 2021.