2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC) | 979-8-3503-2447-1/23/$31.00 ©2023 IEEE | DOI: 10.1109/EMBC40787.2023.10341154

Mobile Communication Log Time Series to Detect Depressive Symptoms

ML Tlachac[1], Miranda Reisch[2], and Michael Heinz[3]

***Abstract*— Major Depressive Disorder (MDD) is highly prevalent and characterized by often debilitating behavioral and cognitive symptoms. MDD is poorly understood, likely due to considerable heterogeneity and self-report-driven symptomatology. While researchers have been exploring the ability of machine learning to screen for MDD, much less attention has been paid to individual symptoms. We posit that understanding the relationship between objective data streams and individual depression symptoms is important for understanding the considerable heterogeneity in MDD. Thus, we conduct a comprehensive comparative study to explore the ability of machine learning to predict nine self-reported depressive symptoms with call and text logs. We created time series from the logs of over 300 participants by aggregating communication attributes– average length, count, or contacts– every 4, 6, 12, or 24 hours. We were most successful predicting movement irregularities with a balanced accuracy of 0.70. Further, we predicted suicidal ideation with a balanced accuracy of 0.67. Outgoing texts proved to be the most useful log type. This study provides valuable insights for future mobile health research aimed at personalizing assessment and intervention for MDD.**

I. INTRODUCTION

Major Depressive Disorder (MDD) is a highly prevalent and burdensome mental disorder [1], [2], characterized by varying groupings of co-occurring symptoms, which include low interest, depressed mood, concentration difficulties, and suicidal thoughts [3]. MDD is highly heterogeneous in its clinical presentation, with over 1000 distinct profiles by one estimate [4]. Clinically, the existence of depressive symptoms are assessed through self-reporting measures, such as depression screening surveys [5]. Unfortunately, depressive symptoms are often debilitating and interfere with help seeking behavior [6]. Further, patients may not recognize or be willing to disclose all symptoms [7]. Thus, the diagnostic construct MDD is poorly understood and often misdiagnosed or under-diagnosed [8]. As such, there is a need for a unobtrusive approach to identify and track depressive symptoms.

Research in the field of applied machine learning has made important contributions to mental disorder screening and diagnosis. Mobile modalities are particularly promising for MDD screening given their ease of collection with prior research using voice recordings [9], environmental audio [10], location data [11], [12], [13], received text content [14], sent text content [15], [16], and communication logs [13], [17], [18], [19], [20]. Most research to date utilizing mobile modalities focuses on MDD screening at the *disorder* level by aggregating individual symptom severity scores, regardless of which symptoms are contributing to the total depression screening score. To date, only location data has been used to detect individual depressive symptoms [21].

We posit the importance of an approach which accounts for individual MDD *symptoms* for three primary reasons. First, such an approach is inline with existing research initiatives, such as the National Institute of Mental Health (NIMH)'s Research Domain Criteria (RDoC) [22], and the Hierarchical Taxonomy of Psychpathology (HiTOP) [23]; these initiatives' aims include an improved understanding of categorically defined mental disorders through a more nuanced, dimensional, and scientifically grounded approach to psychopathology. Secondly, symptom level detection would make possible the direct screening for high-risk stigmatized depressive symptoms [24] such as self harm, regardless of overall depression severity. Early identification of such symptoms would allow for early and targeted intervention. Lastly, to understand the relative effectiveness of models and sensor modalities in predicting MDD symptoms, conducting thorough benchmark tests is essential, as the prediction signal based on composite MDD screening scores may be weakened by symptoms that are not well modeled. Through benchmarking, we can determine the most effective approach for predicting the diverse symptoms of MDD.

Given that social interactions are known to be important for wellbeing [25], we conduct a comparative assessment of the ability to predict depressive symptoms including suicidal ideation with mobile text and call logs. To preserve privacy, we create time series from the log metadata without content by aggregating communication attributes like communication count at certain intervals. We then extract time series features to use as input to machine learning classifiers. Overall, we compare the ability to detect nine depressive symptoms using four types of communication logs, three time series communication attributes, and four machine learning classifiers.

II. DATA & CLASSIFICATION METHODOLOGY

A. Dataset of Text and Call Logs

We use retrospectively harvested SMS text and call logs [18] in the Moodable [26] and EMU [27] datasets. Data was collected between 2017 and 2019 from crowdsourced workers using an Android app. Participants were prompted to complete the PHQ-9 depression screening survey [28] to label the data. Each of the nine questions correspond to a

This research was supported in part by NIH via T32 DA037202.

[1]ML Tlachac is with the Department of Information Systems and Analytics and the Center for Health and Behavioral Sciences, Bryant University, Smithfield, RI 02911 USA mltlachac@bryant.edu

[2]Miranda Reisch is with the Department of Data Science, Worcester Polytechnic Institute (WPI), Worcester, MA 01609, USA mhernandezreisch@wpi.edu

[3]Michael Heinz is with the Departments of Quantitative Biomedical Sciences and Epidemiology at Dartmouth College and the Department of Psychiatry at Dartmouth Health, Hanover, NH 03755 USA michael.v.heinz@dartmouth.edu

TABLE I

OF THE 312 PARTICIPANTS, 182 HAD INCOMING CALLS, 197 HAD OUTGOING CALLS, 290 HAD INCOMING TEXTS, AND 99 HAD OUTGOING TEXTS. WE REPORT PERCENT WITH $Q1 - Q8 \geq 2$ AND $Q9 \geq 1$.

	Call		Text	
Symptom	In	Out	In	Out
Q1:Little interest	45.1%	43.7%	44.5%	49.5%
Q2: Feeling depressed	33.0%	34.0%	37.2%	36.4%
Q3: Trouble sleeping	49.5%	49.2%	52.1%	53.5%
Q4: Feeling tired	54.4%	54.8%	55.9%	61.6%
Q5: Appetite irregularities	44.5%	42.6%	41.7%	46.5%
Q6: Feeling like a failure	37.4%	34.0%	39.3%	40.4%
Q7: Trouble concentrating	39.0%	35.5%	38.6%	46.5%
Q8: Movement irregularities	24.2%	21.3%	24.1%	31.3%
Q9: Self harm thoughts	41.7%	41.1%	44.8%	53.5%

depressive symptom in the DSM-IV. Participants are asked to reflect on the last two weeks when reporting on symptom severity with options "0: Not at all", "1: Several days", "2: More than half the days", and "3: Nearly every day" [28].

To be included in our analysis, we require participants to have at least two incoming texts, two outgoing texts, two minutes of incoming calls, or two minutes of outgoing calls in the two weeks preceding the completion of the PHQ-9. Overall, 312 participants qualified. Incoming texts were shared by the most participants while outgoing texts were shared by the least participants. As is convention [28], we consider a score of at least 2 to be indicative of experiencing depressive symptoms Q1-Q8 and a score of at least 1 to be indicative of experiencing depressive symptom Q9. The number of participants who reported each symptom is in Table I. The most frequent symptom was tiredness (Q4) and the least frequent symptom was movement irregularities (Q8). Notably, Q9 can be considered a measure of suicidal ideation [29]. The subset of participants who shared outgoing text messages reported the highest rate of suicidal ideation. Related research noted that crowdsourced workers have higher rates depression than the general population [10], [19], and the same also seems true of suicidal ideation.

B. Constructing Log Time Series and Extracting Features

We create separate time series for each of the four different log types: incoming texts, outgoing texts, incoming calls, and outgoing calls. For every combination of person and log type, we consider the logs in the two weeks preceding the completion of the PHQ-9. We then group these logs every 4 hours, 6 hours, 12 hours, and 24 hours. We refer to these groupings as the aggregation intervals of the time series.

From the intervals, we calculate three communication attributes: communication count, average communication length, and number of unique contacts. Given the relative scarcity of phone calls, we consider call count to be the number of seconds on a call. In this manner, we form time series of count, average length, and contacts for incoming texts, outgoing texts, incoming calls, and outgoing calls. If a participant shared all four log types, their logs would be represented with 48 time series. Alternatively, if a participant only has one log type, 12 time series would be created.

We use the Time Series Feature Extraction Library [30] to transform the time series into statistical, temporal and spectral features. The time series have different numbers of time steps based on the aggregation interval. Therefore, there were 187, 173, 159, and 152 features extracted respectively for the time series with 4, 6, 12, and 24 hour aggregation intervals. We reduce the dimensionality of the data by creating principal components (PCs) through principal component analysis (PCA) [31]. Since the slope feature produced infinity values, we disregard it. We normalize the features between 0 and 1 prior to applying PCA. These transformations were learned on the training sets and applied to the test sets.

C. Classifiers and Evalaution

We use four common machine learning methods in this exploratory study[1]: Gaussian Naive Bayes (GNB), support vector machine (SVM), logistic regression (LR), and random forest (RF). We train these classifiers with the default parameters [31]. We use between the top one and top five PCs as model input. The training sets are upsampled to balance classes. Given the small number of participants who shared outgoing texts, we use a leave-one-out cross-validation strategy to ensure result robustness. In this form of cross-validation, the test set consists of a single data instance and each is used as the test set once. The number of true positive, false positive, false negative, and true negative predictions are then consolidated. We evaluate models using balanced accuracy, the mean of sensitivity and specificity.

III. RESULTS OF DETECTING DEPRESSIVE SYMPTOMS

The highest balanced accuracy in Tables II-IV is 0.70 with outgoing texts. Surprisingly, this model predicted movement irregularities (Q8). Outgoing text logs was overall the most useful log type, obtaining the highest balanced accuracies for seven symptoms. There is more variation regarding the most useful communication attribute and aggregation interval.

Q1. For Q1, outgoing texts were best for all three communication attributes. The highest balanced accuracy is 0.68 with unique outgoing text contacts aggregated every 24 hours. This GNB with one PC had a sensitivity of 0.73 and a specificity of 0.62. Three other outgoing text models achieved a balanced accuracy of 0.67. Further, outgoing call contacts achieved a balanced accuracy of 0.65 which is better than any count models. We conclude that lack of interest is best predicted with daily number of unique contacts.

Q2. Interestingly, Q2 regarding feeling depressed is the only symptom best predicted with incoming calls. For average incoming call length, SVM with five PCs achieved a balanced accuracy of 0.65, sensitivity of 0.55, and specificity of 0.75. Incoming calls was also best for communication count. Yet, like for Q1, the best modality for unique contacts was outgoing texts with a 24 hour aggregation interval.

Q3. For Q3, outgoing text count aggregated over 12 hours achieved the highest balanced accuracy of 0.66 with a RF and four PCs; the sensitivity was 0.75 and specificity was 0.56.

[1]Code will be available through https://emutivo.wpi.edu/.

TABLE II

THE BALANCED ACCURACY OF THE BEST LEAVE-ONE-OUT CROSS-VALIDATION MODEL CONFIGURATIONS USING *LENGTH* TIME SERIES FEATURES.

	Incoming Call Length				Outgoing Call Length				Incoming Text Length				Outgoing Text Length			
	4hrs	6hrs	12hrs	24hrs	4hrs	6hrs	12hrs	24hrs	4hrs	6hrs	12hrs	24hrs	4hrs	6hrs	12hrs	24hrs
Q1	0.52	0.53	0.59	0.56	0.54	0.53	0.56	0.57	0.55	0.58	0.60	0.56	0.65	0.66	0.66	**0.67**
Q2	0.63	0.63	**0.65**	0.61	0.55	0.54	0.55	0.54	0.50	0.51	0.53	0.50	0.58	0.61	0.59	0.55
Q3	0.58	0.56	0.59	0.54	0.53	0.51	0.51	0.55	0.55	0.54	0.56	0.57	**0.60**	**0.60**	0.58	0.58
Q4	0.56	0.59	0.54	0.54	0.56	0.53	0.52	0.56	0.50	0.54	0.50	0.49	0.56	0.55	0.53	**0.64**
Q5	0.56	0.60	0.55	0.57	0.58	**0.62**	0.60	0.59	0.52	0.53	0.51	0.52	0.57	0.54	0.61	0.58
Q6	0.58	0.54	0.54	0.55	0.54	0.53	0.55	0.54	0.53	0.49	0.52	0.54	0.58	0.59	**0.63**	0.59
Q7	0.54	0.57	0.56	0.59	0.56	0.54	0.58	0.62	0.52	0.54	0.56	0.54	0.60	0.61	**0.63**	0.61
Q8	0.53	0.56	0.58	0.56	0.63	0.54	0.58	0.60	0.57	0.59	0.61	0.58	0.68	**0.70**	0.69	0.65
Q9	0.56	0.54	0.56	0.62	0.56	0.58	0.56	0.58	0.55	0.56	0.61	0.60	0.62	0.63	**0.66**	**0.66**

TABLE III

THE BALANCED ACCURACY OF THE BEST LEAVE-ONE-OUT CROSS-VALIDATION MODEL CONFIGURATIONS USING *COUNT* TIME SERIES FEATURES.

	Incoming Call Count				Outgoing Call Count				Incoming Text Count				Outgoing Text Count			
	4hrs	6hrs	12hrs	24hrs	4hrs	6hrs	12hrs	24hrs	4hrs	6hrs	12hrs	24hrs	4hrs	6hrs	12hrs	24hrs
Q1	0.52	0.57	0.49	0.62	0.54	0.54	0.55	0.59	0.55	0.55	0.54	0.55	0.62	**0.64**	0.63	0.62
Q2	**0.62**	0.61	0.58	0.59	0.55	0.55	0.53	0.58	0.51	0.52	0.52	0.56	0.57	0.58	0.60	0.55
Q3	0.58	0.63	0.56	0.58	0.53	0.56	0.56	0.56	0.56	0.54	0.56	0.55	0.57	0.60	**0.66**	0.59
Q4	0.58	0.58	0.55	0.51	0.53	0.55	0.56	0.58	0.54	0.54	0.54	0.53	0.60	0.63	**0.65**	0.64
Q5	0.55	0.57	0.54	0.54	0.58	**0.61**	0.60	0.58	0.49	0.52	0.52	0.53	0.59	0.56	0.51	0.58
Q6	0.56	**0.58**	0.56	0.55	0.56	0.54	0.56	0.56	0.52	0.57	0.55	0.54	0.56	0.56	0.55	0.55
Q7	0.53	0.54	0.54	0.57	0.57	0.56	0.57	0.57	0.52	0.54	0.55	0.54	0.56	0.56	**0.61**	0.59
Q8	0.54	0.52	0.52	0.61	0.56	0.57	0.56	0.61	0.59	0.62	0.61	0.60	0.62	**0.67**	**0.67**	0.66
Q9	0.58	0.60	0.55	0.59	0.53	0.53	0.52	0.56	0.58	0.55	0.56	0.56	0.63	0.61	0.59	**0.65**

TABLE IV

THE BALANCED ACCURACY OF THE BEST LEAVE-ONE-OUT CROSS-VALIDATION MODEL CONFIGURATIONS USING *CONTACT* TIME SERIES FEATURES.

	Incoming Call Contact				Outgoing Call Contact				Incoming Text Contact				Outgoing Text Contact			
	4hrs	6hrs	12hrs	24hrs	4hrs	6hrs	12hrs	24hrs	4hrs	6hrs	12hrs	24hrs	4hrs	6hrs	12hrs	24hrs
Q1	0.61	0.61	0.60	0.62	0.60	0.65	0.65	0.63	0.55	0.56	0.57	0.62	0.67	0.65	0.67	**0.68**
Q2	0.60	0.59	0.60	0.60	0.60	0.56	0.58	0.57	0.49	0.50	0.51	0.52	0.60	0.54	0.56	**0.63**
Q3	**0.61**	0.58	0.57	0.59	0.55	0.58	0.59	0.56	0.54	0.51	0.52	0.52	0.56	0.58	0.58	0.56
Q4	0.54	0.53	0.50	0.55	0.56	0.59	**0.60**	0.57	0.53	0.54	0.54	0.51	0.53	0.54	0.58	0.58
Q5	0.56	0.55	0.54	0.60	0.58	0.59	0.59	0.58	0.57	0.56	0.53	0.55	0.57	0.59	0.60	**0.61**
Q6	0.58	0.59	0.56	0.59	**0.60**	0.58	0.57	0.57	0.53	0.52	0.54	0.55	0.56	**0.60**	0.56	0.57
Q7	0.58	0.57	0.57	0.59	0.58	0.58	0.57	0.58	0.54	0.57	0.54	0.55	0.61	0.59	0.57	**0.63**
Q8	0.59	0.54	0.58	0.57	0.58	0.58	0.58	0.60	0.57	0.61	0.55	0.58	0.65	**0.68**	0.67	0.67
Q9	0.59	0.60	0.57	0.58	0.59	0.60	0.57	0.58	0.56	0.55	0.60	0.57	0.64	0.64	0.64	**0.67**

Apart from incoming call count aggregated over 6 hours, no other model achieved a balanced accuracy above 0.6. Communication count was thus most indicative of trouble sleeping.

Q4. For Q4, the two highest balanced accuracies of 0.65 and 0.64 were both achieved with outgoing texts. While 24 hours was much better than 12 hours for average outgoing text length, both aggregation intervals were successful for outgoing text count. The best classifier was a SVM with three PCs; it achieved a sensitivity of 0.51 and specificity of 0.79. Unique contacts were not as helpful to detect tiredness.

Q5. Appetite irregularities is understandably the most challenging symptom to detect with logs. It is the only symptom best predicted with outgoing calls. All communication attributes had similar screening abilities with balanced accuracies between 0.61 and 0.62. Overall, 6 hour aggregation intervals was most useful. The best model, a RF with one PC, has a sensitivity of 0.51 and a specificity of 0.73.

Q6. Outgoing text length aggregated over 12 hours had the highest balanced accuracy of 0.63 for Q6. LR with five PCs has a sensitivity of 0.73 and a specificity of 0.54. While count features were particularly unhelpful for predicting feelings of failure, both outgoing calls and texts had the same balanced accuracy of 0.60 for number of unique contacts.

Q7. For Q7, two GNB on outgoing texts had the highest balanced accuracy of 0.63. With three PCs, average text length aggregated over 12 hours yields a sensitivity of 0.76 and a specificity of 0.49. With two PCs, daily unique outgoing text contacts yields a sensitivity of 0.61 and specificity of 0.64. Daily outgoing call length was also predictive of trouble concentrating with a balanced accuracy of 0.62.

Q8. As mentioned, predicting Q8 was most successful. All three communication attributes performed well with outgoing texts aggregated every six hours. The best model, a RF with two PCs, used average length of outgoing texts With a sensitivity of 0.55 and a specificity of 0.85, it is more useful for eliminating participants without movement irregularities.

Q9. For all three communication attributes, daily outgoing texts had the highest balanced accuracy when predicting thoughts of self harm. Unique contacts was most successful with a balanced accuracy of 0.67, sensitivity of 0.64, and

specificity of 0.70. These results were from a LR with two PCs, though a GNB with one PC performed similarly.

IV. Discussion with Related & Future Work

The objective of our comparative study was to explore the potential of using passively collected call and text log metadata to predict symptom-level depression scores, thus aligning with the RDoC [22] and the HiTOP [23] frameworks which advocate for a dimensional approach to mental disorders. We focused on identifying the relationship between specific metadata sensor streams and depressive symptoms. Overall, outgoing texts were the most predictive log type. The aggregation interval to use was evident for Q1, Q4, Q5, Q8, and Q9. Likewise, the communication attribute to (not) use was evident for Q1, Q3, Q4, Q5, and Q6. For example, daily number of unique contacts was most predictive for Q1.

When screening for moderate depression, log time series features achieved a balanced accuracy of 0.66 [18]. We achieved higher balanced accuracy when predicting little interest (Q1), movement irregularities (Q8), and thoughts of self harm (Q9). Prior research [26], [29], [9] has also predicted suicidal ideation (Q9) with large variability in balanced accuracies from 0.62 with multimodal mobile features [26] to 0.81 with text content features [29]. Unfortunately, these previously explored modalities have privacy concerns.

The study [21] that used location data to predict depressive symptoms had insufficient participants who responded to Q9 in the affirmative for modeling. Likewise, due to lack of reported symptoms, it predicted $Q1 - Q8 \geq 1$. The highest balanced accuracies were 0.76 for Q1, 0.76 for Q2, 0.70 for Q3, 0.70 for Q4, 0.80 for Q5, 0.79 for Q6, 0.70 for Q7, and 0.66 for Q8 [21]. Interestingly, their location data was most successful at predicting Q5 and least successful at predicting Q8 whereas our communication logs were most successful at predicting Q8 and least successful at predicting Q5.

Like related digital phenotype research [21], [9], the number of participants is a limitation. As we opted to retain as many participants as possible, participants shared different logs types and quantities. We also assumed participants used their personal phones. While changing communication trends [32] could be considered a limitation, future research can still apply insights gleaned from our finding when using communication logs from other platforms. Such research could combine logs from multiple sources into multimodal classifiers to identify and track depressive symptoms.

ACKNOWLEDGMENT

We thank Veronica Melican, Elke Rundensteiner, Ermal Toto, and prior Emutivo teams at WPI for data contributions.

References

[1] S. Avenevoli *et al.*, "Major depression in the national comorbidity survey–adolescent supplement: Prevalence, correlates, and treatment," *J Am Acad Child Adolesc Psychiatry*, vol. 54, no. 1, pp. 37–44, 2015.

[2] D. Proudman, P. Greenberg, and D. Nellesen, "The growing burden of major depressive disorders (MDD): Implications for researchers and policy makers," *PharmacoEconomics*, vol. 39 (6), pp. 619–625, 2021.

[3] American Psychiatric Association, *Diagnostic and Statistical Manual of Mental Disorders, 5th Ed. (DSM-5)*. Am. Psych. Publishing, 2013.

[4] E. I. Fried and R. M. Nesse, "Depression is not a consistent syndrome: An investigation of unique symptom patterns in the STAR*D study," *Journal of Affective Disorders*, vol. 172, pp. 96–102, 2015.

[5] M. Savoy and D. O'Gurek, "Screening your adult patients for depression," *Family practice management*, vol. 23, no. 2, pp. 16–20, 2016.

[6] K. Demyttenaere, A. Bonnewyn *et al.*, "Comorbid painful physical symptoms and depression: prevalence, work loss, and help seeking," *Journal of affective disorders*, vol. 92, no. 2-3, pp. 185–193, 2006.

[7] R. Epstein, P. Duberstein *et al.*, ""I didn't know what was wrong:" how people with undiagnosed depression recognize, name and explain their distress," *J. Gen. Intern. Med.*, vol. 25 (9), pp. 954–61, 2010.

[8] E. J. Pérez-Stable, J. Miranda, R. F. Muñoz, and Y.-W. Ying, "Depression in medical outpatients: underrecognition and misdiagnosis," *Archives of Internal Medicine*, vol. 150, no. 5, pp. 1083–1088, 1990.

[9] M. L. Tlachac, R. Flores, E. Toto, and E. Rundensteiner, "Early mental health uncovering with short scripted and unscripted voice recordings," *Deep Learning Applications*, vol. 4, 2022.

[10] D. Di Matteo, K. Fotinos *et al.*, "The relationship between smartphone-recorded environmental audio and symptomatology of anxiety and depression: exploratory study," *JMIR Form. Res.*, vol. 4, no. 8, 2020.

[11] A. Farhan *et al.*, "Behavior vs. introspection: refining prediction of clinical depression via smartphone sensing data," in *IEEE WH*, 2016.

[12] S. Ware *et al.*, "Large-scale automatic depression screening using meta-data from wifi infrastructure," *ACM IMWUT*, vol. 2, no. 4, 2018.

[13] P. Chikersal, A. Doryab *et al.*, "Detecting depression and predicting its onset using longitudinal symptoms captured by passive sensing: a machine learning approach with robust feature selection," *ACM TOCHI*, vol. 28, no. 1, pp. 1–41, 2021.

[14] M. L. Tlachac *et al.*, "You're making me depressed: Leveraging texts from contact subsets to predict depression," in *IEEE BHI*, 2019.

[15] M. L. Tlachac and E. Rundensteiner, "Screening for depression with retrospectively harvested private versus public text," *IEEE J-BHI*, vol. 24, no. 11, pp. 3326–32, 2020.

[16] M. L. Tlachac *et al.*, "Automated construction of lexicons to improve depression screening with text messages," *IEEE J-BHI*, 2022.

[17] M. L. Tlachac and E. A. Rundensteiner, "Depression screening from text message reply latency," in *42nd IEEE EMBC*, 2020, pp. 5490–3.

[18] M. L. Tlachac, V. Melican, and et al., "Mobile depression screening with time series of text logs and call logs," in *IEEE BHI*, 2021.

[19] M. L. Tlachac, R. Flores *et al.*, "DepreST-CAT: Retrospective smartphone call and text logs collected during the covid-19 pandemic to screen for mental illnesses," *ACM IMWUT*, vol. 6, no. 2, 2022.

[20] M. L. Tlachac and S. S. Ogden, "Left on read: Reply latency for anxiety & depression screening," *ACM UbiComp*, 2022.

[21] S. Ware, C. Yue *et al.*, "Predicting depressive symptoms using smartphone data," *Smart Health*, vol. 15, 2020.

[22] B. N. Cuthbert, "The RDoC framework: facilitating transition from ICD/DSM to dimensional approaches that integrate neuroscience and psychopathology," *World Psychiatry*, vol. 13, no. 1, pp. 28–35, 2014.

[23] C. C. Conway, M. K. Forbes, and S. C. South, "A hierarchical taxonomy of psychopathology (hitop) primer for mental health researchers," *Clinical Psychological Science*, vol. 10, no. 2, pp. 236–258, 2022.

[24] L. Martin, H. Neighbors, and D. Griffith, "The experience of symptoms of depression in men vs women: analysis of the national comorbidity survey replication," *JAMA psychiatry*, vol. 70 (10), pp. 1100–6, 2013.

[25] S. Cohen, "Social relationships and health," *American Psychologist*, vol. 59, no. 8, 2004.

[26] A. Dogrucu, A. Perucic *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*, vol. 17, pp. 100–18, 2020.

[27] M. L. Tlachac, E. Toto *et al.*, "Emu: Early mental health uncovering framework and dataset," in *20th IEEE ICMLA*, 2021, pp. 1311–18.

[28] 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.

[29] M. L. Tlachac, K. Dixon-Gordon, and E. Rundensteiner, "Screening for suicidal ideation with text messages," in *IEEE BHI*, 2021.

[30] M. Barandas, D. Folgado *et al.*, "Tsfel," *SoftwareX*, vol. 11, 2020.

[31] F. Pedregosa, G. Varoquaux *et al.*, "Scikit-learn: Machine learning in python," *J of Machine Learning Research*, vol. 12, pp. 2825–30, 2011.

[32] B. Wetzel, R. Pryss *et al.*, ""how come you don't call me?" smartphone communication app usage as an indicator of loneliness and social well-being across the adult lifespan during the covid-19 pandemic," *Int J of Environmental Research and Public Health*, vol. 18, no. 12, 2021.