Journal Pre-proof

Detecting Major Depressive Disorder Presence Using Passively-Collected Wearable Movement Data in a Nationally-Representative Sample

George D. Price , Michael V. Heinz , Amanda C. Collins , Nicholas C. Jacobson

PII: S0165-1781(23)00643-1
DOI: https://doi.org/10.1016/j.psychres.2023.115693
Reference: PSY 115693

To appear in: *Psychiatry Research*

Received date: 18 August 2023
Revised date: 21 December 2023
Accepted date: 24 December 2023

Please cite this article as: George D. Price , Michael V. Heinz , Amanda C. Collins , Nicholas C. Jacobson , Detecting Major Depressive Disorder Presence Using Passively-Collected Wearable Movement Data in a Nationally-Representative Sample, *Psychiatry Research* (2023), doi: https://doi.org/10.1016/j.psychres.2023.115693

Detecting Major Depressive Disorder Presence Using Passively-Collected Wearable Movement Data in a Nationally-Representative Sample

George D. Price[a,b], Michael V. Heinz[a,c], Amanda C. Collins[a,d], Nicholas C. Jacobson[a,b,c,d]

[a]Center for Technology and Behavioral Health, Geisel School of Medicine, Dartmouth College, Lebanon, NH, United States

[b]Quantitative Biomedical Sciences Program, Dartmouth College, Lebanon, NH, United States

[c]Department of Psychiatry, Geisel School of Medicine, Dartmouth College, Lebanon, NH, United States

[d]Department of Biomedical Data Science, Geisel School of Medicine, Dartmouth College, Lebanon, NH, United States

Author Note

George D. Price https://orcid.org/0000-0002-9164-4973

Michael V. Heinz https://orcid.org/0000-0003-0866-0508

Amanda C. Collins https://orcid.org/0000-0002-8258-2272

Nicholas C. Jacobson https://orcid.org/0000-0002-8832-4741

***Correspondence:**

George Price, Dartmouth College, 46 Centerra Parkway Suite 300, Office # 336S, Lebanon, NH 03766.

Email: George.Price.GR@dartmouth.edu

Word Count: 4001/5000

Abstract

Major Depressive Disorder (MDD) is a heterogeneous disorder, resulting in challenges with early detection. However, changes in sleep and movement patterns may help improve detection. Thus, this study aimed to explore the utility of wrist-worn actigraphy data in combination with machine learning (ML) and deep learning techniques to detect MDD using a commonly used screening method: Patient Health Questionnaire-9 (PHQ-9). Participants (N = 8,378; MDD Screening = 766 participants) completed the and wore Actigraph GT3X+ for one week as part of the National Health and Nutrition Examination Survey (NHANES). Leveraging minute-level, actigraphy data, we evaluated the efficacy of two commonly used ML approaches and identified actigraphy-derived biomarkers indicative of MDD. We employed two ML modeling strategies: (1) a traditional ML approach with theory-driven feature derivation, and (2) a deep learning Convolutional Neural Network (CNN) approach, coupled with gramian angular field transformation. Findings revealed movement-related features to be the most influential in the traditional ML approach and nighttime movement to be the most influential in the CNN approach for detecting MDD. Using a large, nationally-representative sample, this study highlights the potential of using passively-collected, actigraphy data for understanding MDD to better improve diagnosing and treating MDD.

Keywords: major depressive disorder; depression presence; movement behavior; sedentary behavior; sleep behavior

1 Introduction

Major depressive disorder (MDD) is the leading cause of disability worldwide (Reddy, 2010), and the prevalence of MDD has steadily increased in the United States over the past two decades (Case & Deaton, 2015; Mojtabai et al., 2016). MDD is heterogeneous in clinical presentation, characterized by symptoms including low mood, diminished pleasure or interest, appetite/weight disturbances, sleep problems, psychomotor difficulties, fatigue, negative self-views, concentration difficulties, and suicidal ideation (Kennedy, 2022; Maurice-Tison et al., 1998; Riemann et al., 2001). Individuals with MDD experience distress and significant impairments in functional, social, and occupational domains (Park & Jung, 2019). These impairments may include an inability to work, complete daily activities (e.g., bathing), or be mobile (e.g., exercising; Kessler et al., 2003).

Given the high prevalence and significant functional impairments associated with MDD, the low detection and diagnosis rates are a significant public health concern (Manea et al., 2015). Specifically, only 4.2% of adults without an existing diagnosis of MDD are screened for MDD at their medical appointments, and only 47% of these visits result in a new diagnosis of MDD (Akincigil & Matthews, 2017). The nonspecific nature of many MDD symptoms and considerable overlap with physical, nonspecific symptoms (e.g., fatigue) may contribute to these low screening and detection rates. Indeed, physical symptoms are often misattributed to other medical issues (Hallyburton & Allison-Jones, 2023), leading to missed opportunities for depression screening and diagnosis. To improve detection of MDD, it is important to identify additional biomarkers to supplement self-report questionnaires, as undiagnosed and untreated MDD can lead to long-term disability (Ghio et al., 2015; Hung et al., 2015).

Investigating sedentary, movement, and sleep behaviors may improve the detection of MDD. Meta-analytic findings suggest that sedentary behavior (e.g., working at a computer, watching TV) is associated with a depression; however, only 14% of the included studies utilized objective measures of sedentary behavior, thus their findings are largely based off of self-report measures (Zhai et al., 2015). Recent work utilizing objective measures indicates that depressed individuals spend more time engaged in sedentary behaviors and less time daily engaged in light-intensity physical activity (e.g., walking) and moderate-to-vigorous physical activity (e.g., exercise; Gianfredi et al., 2022). Relatedly, sleep disturbances (either insomnia or hypersomnia) are common in depression. There is a bidirectional relationship between sleep difficulties and depression with sleep difficulties increasing the risk for depression and vice versa.(Franzen & Buysse, 2008) While prior research has examined the relationship between sedentary behavior, movement, and sleep with depression (Bailey et al., 2018; Camacho et al., 1991; Franzen & Buysse, 2008; Mammen & Faulkner, 2013; Marques et al., 2020; Schuch & Stubbs, 2019; Vallance et al., 2011), the extent to which these constructs can characterize or detect MDD has been limited due to response bias and reporting frequency (Ameko et al., 2018; Onnela & Rauch, 2016).

Actigraphy, specifically with the use of a triaxial accelerometer (e.g., Actigraph GT3X+; Skotte et al., 2014), is a more reliable method of collecting activity data compared to self-report questionnaires (Patterson et al., 1993). Furthermore, actigraphy provides a more accurate detection of sleep behavior and duration compared to subjective measures of patient report in general populations and primary-mood disorder clinical populations (Biddle et al., 2015). Actigraphy offers an objective, unobtrusive method of gathering long-term behavioral information in a naturalistic setting (DeMasi & Recht, 2017; Nguyen et al., 2017), which has shown promise in early detection of MDD (Opoku Asare et al., 2021), thus providing an opportunity for early detection that might not otherwise be captured within self-report questionnaires. Naturalistic actigraphy data can thus be leveraged in conjunction with analytic approaches, such as machine learning, to investigate how activity patterns can detect MDD.

Machine learning has been previously used to detect MDD from actigraphy data (Biddle et al., 2015; Mammen & Faulkner, 2013; Patterson et al., 1993), as well as construct personalized models to more accurately predict individual-level treatment response (Mellem et al., 2020; S. Wang et al., 2019). In modeling dense sequence data, such as that gathered from actigraphic sensors, two general approaches may be utilized. One, our *theory-organized approach*, relies on existing knowledge of MDD to inform the manual derivation of features that reflect summative statistical measures of the time-series data (e.g., average daily movement). The second approach, our *self-organizing approach*, utilizes deep learning models (e.g., convolutional neural networks) with the raw, minute-level actigraphy data to derive latent features, agnostic to existing domain knowledge. The ability of theory-organized and self-organizing approaches to handle data that are collected non-invasively, in real time (Kushki et al., 2019; Meinlschmidt et al., 2020), allows for quantification of an individual's behavioral patterns using technology, defined as *digital phenotyping* (Nemesure et al., 2021). The construction and analysis of digital biomarkers (Hudson & Collins, 2017) for persons with MDD based on their digital phenotypes may serve as a promising approach to better understand their behavioral patterns, and subsequently provide the opportunity for improved MDD detection.

The purpose of the current study is to develop and implement computational methods for identifying digital biomarkers to improve detection and understanding of MDD using passively-collected actigraphy data. The study will implement, independently and in parallel, theory-organized and self-organizing machine learning modeling frameworks using strictly passively-collected actigraphy data. Moreover, our work will interrogate the respective strengths of theory and self-organizing-driven frameworks in the detection of MDD in a large, nationally-representative sample. In doing so, not only will framework-specific MDD detection be assessed, but influential theory-organized and self-organizing features will be identified, thus aiding in our understanding of how to detect MDD, and what digital biomarkers are most influential for detection. Further our work will allow for a direct comparison of a traditional machine learning approach operating on theory derived features and a deep learning approach operating on minimally processed time series data. This comparison is important given the relative trade offs of a deep learning approach compared to a traditional machine learning approach (Janiesch et al., 2021). Given that a self-organizing approaches has been shown to allow for enhanced mental health disorder detection in high-dimensional data, (Gao et al., 2018; Kim et al., 2016) we hypothesize that the self-organizing approach will be more effective in detecting MDD presence compared to the theory-organized approach.

2 Methods

2.1 Study Sample

The present study uses participant data from the National Health and Nutrition Examination Survey (NHANES; (Centers for Disease Control and Prevention, 2020a, 2020b). NHANES, a product of the National Center for Health Statistics (NCHS), is a large, nationally-representative study aimed at assessing the health and nutritional status of individuals in the United States via interviews and physical examination information. Ethics approval was received from the NCHS Research Ethics Review Board (Protocol #2011–2017).

Table 1

Baseline Demographics.

Characteristic	Participants, No. (%) MDD (n = 766)	No MDD (n = 7,612)	*P*-value
Mean (SD) age, years	48.61 (16.95)	47.48 (18.69)	0.11
Sex			
Female	499 (65.14)	3791 (49.8)	< 0.001
Male	267 (34.86)	3821 (50.2)	
Race and Ethnicity			
Non-Hispanic			
Black	174 (22.71)	1820 (23.91)	< 0.001
White	327 (42.69)	3072 (40.36)	
Asian	29 (3.79)	907 (11.92)	
Mexican American	85 (11.10)	908 (11.93)	
Other Hispanic	112 (14.62)	682 (8.96)	
Other Race (Including Multi-Racial)	39 (5.09)	223 (2.93)	

Note. Individuals with a total PHQ-9 ≥ 10 were included in the MDD Cohort. A t-test was performed for mean Age, and the χ^2 test was performed for Sex and Race and Ethnicity demographic variables.

2.2 Data Collection and Study Measures

Our sample comprised NHANES participants from collection cycles 2011-2012 and 2013-2014 (*N* = 8,378; (Centers for Disease Control and Prevention, 2005), including wrist-worn actigraphy information and depression scores (PHQ-9) (See Table 1). Screening for potential MDD presence was defined by a PHQ-9 composite score ≥ 10, a threshold which has been validated as an acceptable cut point for MDD detection in multiple studies (Kroenke et al., 2001b, 2001a) and meta analysis (Moriarty et al., 2015).

Actigraphy data was collected via an Actigraph GT3X+, which was provided to participants during their session in the Mobile Examination Center (MEC) and programmed to begin detecting and recording acceleration information at the end of the participant's MEC session. The device was worn on participants' non-dominant hands for atleast seven full days (midnight to midnight; (Centers for Disease Control and Prevention, 2020a, 2020b). External quality control of the actigraphy data was completed by contractors at Northeastern University under the direction of collaborators from the National Center for Health Statistics and the National Cancer Institute to consider properties of the raw acceleration measures that were unlikely to be the results of human movement, such as spikes in values, impossible values, or long periods of irregular minimum or maximum values (Centers for Disease Control and Prevention, 2022). Following quality control, the raw acceleration measures were converted to Monitor Independent Movement Summary (MIMS) units, an open-source method of representing accelerometry information agnostic to the device used for data collection.

The present analyses used the minute-level triaxial MIMS value, which reflects the sum of the individual MIMS measurements obtained from the x-, y-, and z- axes, respectively. MIMS-units offer a single summary metric which is designed to maximize informative signals while simultaneously filtering environmental and movement artifacts (John et al., 2019) and provide a method for comparing device-based measures of physical activity across accelerometer devices (Belcher et al., 2021). As absolute time was not provided in relation to actigraphy data collection, relative device collection start time and total collection time were standardized across participants by using the timestamp corresponding to the first minute of data collection to infer midnight of the first day. Participants data was then truncated to include the subsequent seven days (midnight to midnight), resulting in 10,080 minutes of consecutive actigraphy information (Trost et al., 2005), an established, age-agnostic, observation period for movement-related behavior.

2.3 Data Preprocessing and Feature Engineering

The actigraphy data was pre-processed and analyzed using Python (v 3.9; Van Rossum & Drake, 2009) in two independent, parallel pipelines to accommodate the theory-organized and self-organizing machine learning approaches.

2.3.1 Theory-Organized Approach

The seven days of minute-level actigraphy data were labeled with the day of the week, based on the provided first and last days of data collection according to the NHANES data documentation.(Centers for Disease Control and Prevention, 2020a, 2020b) Subsequently, we derived statistical summative metrics for movement, sedentary and sleep behaviors at the weekly, weekday, and weekend level to address the potential for differential behavioral patterns based on periods of the week (Figure 1, Panel B, Left). Further information regarding feature derivation is provided in the Supplementary Material, and a complete list of the derived features is provided in Supplementary Table 1.

2.3.2 Self-Organizing Approach

The 1D minute-level actigraphy data was transformed to a 2D image using a Gramian angular field (GAF) transformation. GAF transformation allows for the encoding of time-series data in 2D image form, while preserving temporal correlation (Wang & Oates, 2015), which can be analyzed utilizing traditional Convolutional Neural Networks, developed for image classification (Figure 1, Panel B, Right).

2.4 Machine Learning Modeling

2.4.1 Theory-Organized Approach

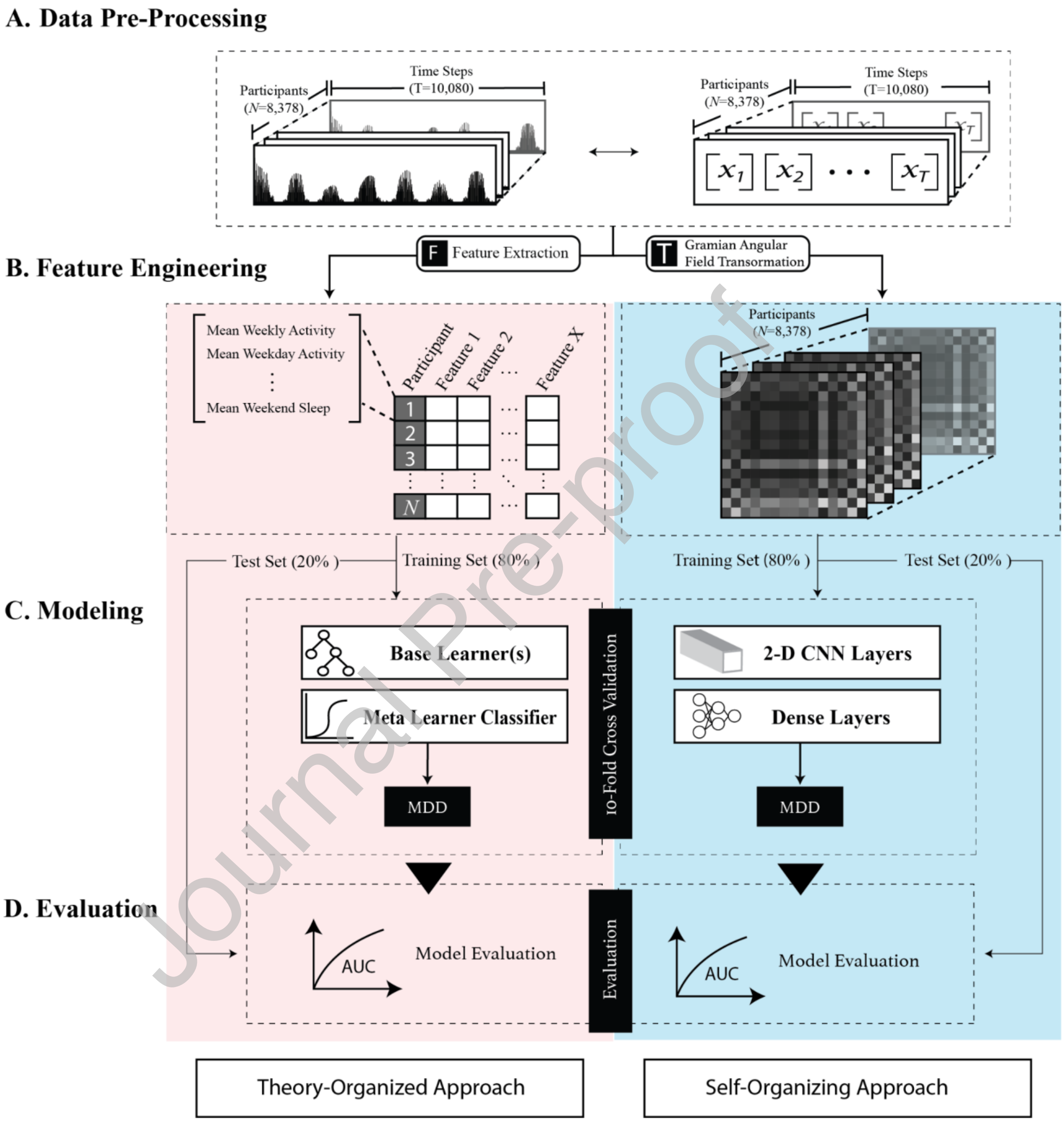
Using the Sklearn package for Python, we split the data randomly into a training-validation set (80%) and a test set (20%). Given the considerable class imbalance, we implemented stratification based on outcome class in order to preserve equal class prevalence between the training-validation and test sets. Using the training-validation set, we implemented a 10-fold cross validation framework in order to gain a robust measure of model performance varying data splits and hyperparameters. After model fitting with the training-validation set, we utilized the fully held out test set (20%) to assess the model's generalizability and predictive performance on unseen data (Figure 1, Panel C, Left; Lekkas et al., 2021). A logistic regression model was used to assess the initial signal in the data, followed by a stacked ensemble approach, which leverages algorithmically distinct models to improve predictive performance compared to individual model selection, and has been shown to consistently outperform base algorithms in mental health disorder-related outcomes (Supplementary Table 2; Ozcift & Gulten, 2011; Tao et al., 2021). Further, per the NHANES Analytic Guidelines, the Mobile Examination Center individual sample weights were incorporated into the modeling pipeline,(Centers for Disease Control and Prevention, 2005) and the outcome class weights were balanced (Johnson & Khoshgoftaar, 2019). Model performance was reported for the Logistic Regression and Stacked Ensemble models as: Area Under the ROC Curve (AUC), sensitivity, and specificity for both the validation set(s) and held-out test set (See Table 2).

2.4.2 Self-Organizing Approach

The cross-validation schema mirrored that outlined in Methods section 2.4.1 (Figure 1, Panel C, Right). To benchmark the deep learning approach, we began with a long short-term memory (LSTM) network; a type of recurrent neural network, which can account for long and short-term temporal dependencies in time-series data (Hochreiter & Schmidhuber, 1997; Pham, 2021) and has been used successfully with actigraphic data to detect mental health-related outcomes (Heinz et al., 2022). The LSTM model was followed with a modified architecture of the robust “AlexNet,” described by Krizhevsky et al. (2017). As referenced in 2.4.1, sample weights and class weights were incorporated into the modeling pipeline. To compute metrics for the held-out test set, we implemented a weighting scheme such that the weight of each model’s predictions was proportional to that model’s performance of the validation set; models with a validation performance with AUC < 0.5 were weighted 0. Subsequently, performance metrics were reported in parallel with those outlined in Methods section 2.4.1 (Table 2). Specific modeling architecture alterations for the LSTM and AlexNet models are outlined in the Supplementary Material.

Figure 1

Modeling Pipeline



Note. *An overview of the modeling pipeline for both theory-organized (traditional machine learning; red) and self-organizing (deep learning; blue) approaches, highlight the distinct but parallel methods of data pre-processing, feature engineering, and modeling.*

2.5 Model Introspection

Given the model(s) complexity, we utilized computational techniques for model introspection – that is, to better understand the magnitude and directionality of the input features most important for model inference.

2.5.1 Theory-Organized Approach

SHapley Additive exPlanations (SHAP) was implemented to assess magnitude and directionality of the five most influential features on the stacked ensemble models predictions. SHAP iteratively perturbs the features of a model and assesses how this affects the model prediction,(Lundberg & Lee, 2017) thereby determining relative feature importance (Figure 2 feature order), as well as the marginal contribution of each feature to an individual's model prediction (dot positioning on x-axis, Figure 2).

2.5.2 Self-organizing Approach

To maintain the temporal density of the input data in the self-organizing approach, SHAP was used quantify the relative importance of a given time point (minute) averaged across participants and days (Heinz et al., 2022). The resulting SHAP values were then averaged over a 60-minute rolling window and mapped to the corresponding MIMS-value for the outcome groups (Figure 4).

3 Results

3.1 Theory-organized Approach

3.1.1 Theory-organized: Model Performance

A logistic regression model was implemented as a benchmark in our theory-organized approach. The logistic regression model performed poorly (AUC_{test} = 0.55, $AUC_{validation}$ = 0.55 ±

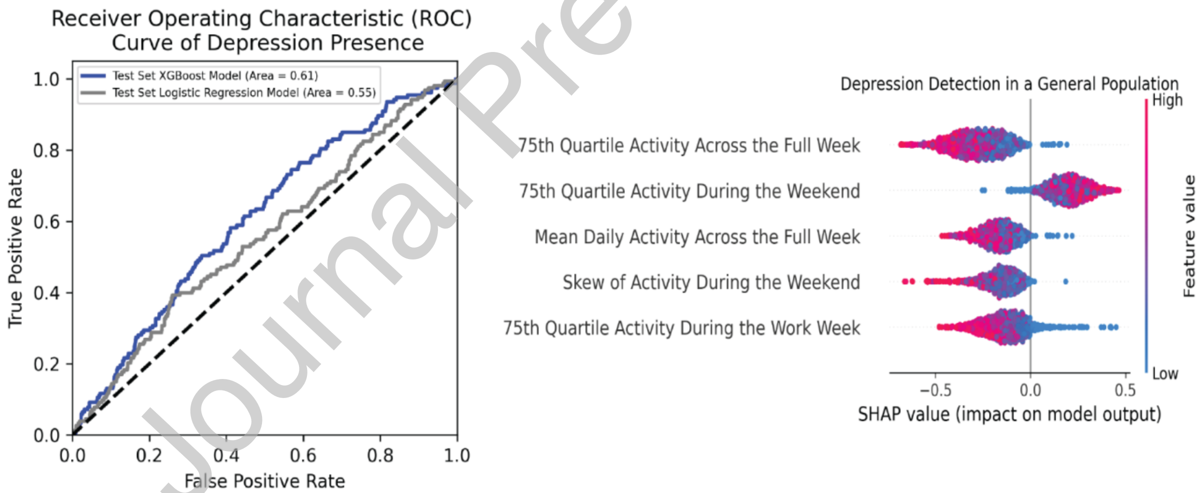
0.04); however, the stacked ensemble model demonstrated moderate performance (AUC_{test} = 0.61, $AUC_{validation}$ = 0.61 ± 0.03) in detecting MDD presence.

3.1.2 Theory-organized: Model Introspection

Using SHAP (Methods section 2.5.1), we found the following features to be the most influential to the model's predictions for MDD presence: (1) Lower high-intensity activity across the full week (represented by 75th Quartile); (2) Higher high-intensity activity during the weekend (represented by 75th quartile); (3) Lower average activity across the full week; (4) A negative skew (left-skew) of weekend activity; (5) Lower high-intensity activity during the work week (represented by 75th quartile).

Figure 2

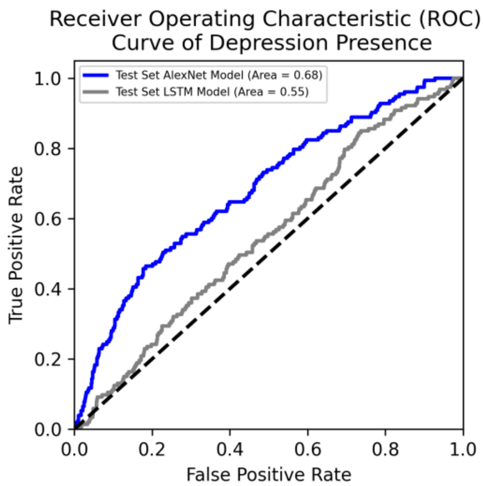
Theory-Organized Model



Note. (Left) The test set AUC of the theory-organized logistic regression and stacked ensemble models against the line of identity. (Right) The plot displays the top five most influential features for the stacked ensemble models predictions. The color of the dot reflects the value of the feature, and a dot's position on the x-axis relative to the origin reflects the individual value's impact on the model prediction.

Figure 3

Self-Organizing Model



Note. The test set AUC of the self-organizing LSTM and AlexNet models against the line of identity.

3.2 Self-organizing Approach

3.2.1 Self-organizing: General Population Model Performance

An LSTM model was implemented as a benchmark model, and performed only marginally above chance (AUC_{test} = 0.55, $AUC_{validation}$ = 0.52 ± 0.03). However, the modified "AlexNet" model showed the best performance across all approaches (AUC_{test} = 0.68, $AUC_{validation}$ = 0.63 ± 0.03) in detecting MDD in a general population.

Table 2

Modeling Approaches

Modeling Approach	Test Set				Validation Set(s)			
	AUC	Optimal Cut Point	Sensitivity	Specificity	AUC	Optimal Cut Point	Sensitivity	Specificity
Theory-organized: Logistic Regression	0.55	0.10	0.39	0.74	0.58 ± 0.04	0.09 ± 0.01	0.69 ± 0.14	0.48 ± 0.15
Theory-organized: Stacked Ensemble	0.61	0.44	0.74	0.43	0.61 ± 0.03	0.49 ± 0.02	0.68 ± 0.10	0.51 ± 0.11
Self-Organizing: LSTM	0.55	0.33	0.84	0.26	0.52 ± 0.03	0.37 ± 0.11	0.44 ± 0.25	0.59 ± 0.27
Self-Organizing: AlexNet with GAF transformation	0.68	0.45	0.45	0.82	0.63 ± 0.03	0.40 ± 0.16	0.57 ± 0.13	0.66 ± 0.11

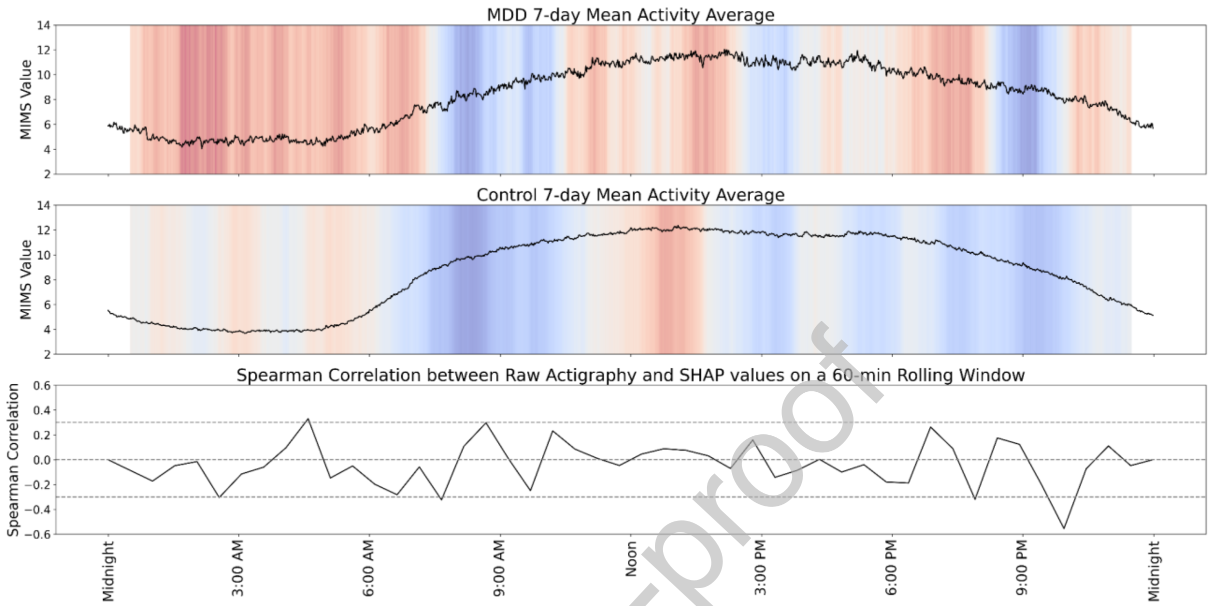
Note. Model performance of the theory-organized and self-organizing machine learning approaches for the validation sets and held-out test set, reported as area under the receiver operating characteristic curve (AUC). LSTM = Long-Short Term Memory, GAF = Gramian Angular Field

3.2.2 Self-organizing: General Population Model Introspection

On qualitative assessment, we note less defined boundaries between those times of typical sleep and wakefulness for individuals with MDD. In particular, we observe a more gradual movement increase in the morning (6AM-11AM) and a more gradual movement taper in the evening (6PM-12AM) compared to those without MDD. We used SHAP to interrogate the modified AlexNet modeling pipeline (Figure 4), and found target diurnal areas highly influential in discriminating between those with and without MDD. In particular, we note areas of high influence in detecting MDD overnight (12AM-6AM), whereas the mid-morning (7AM-10AM) and evening (8PM-10PM) prove influential in detecting individuals without MDD.

Figure 4

SHapley Additive exPlanations (SHAP) Results



Note. *(Top) A 60-minute rolling average MIMS value across participants in the MDD Presence group mapped to a 24-hour cycle. The background color reflects a 60-minute rolling average SHAP value (red = high SHAP value [more predictive of MDD], blue = low SHAP value [more predictive of Non-MDD]). (Middle) A 60-minute rolling average MIMS value across participants in the No MDD Presence group mapped to a 24-hour cycle. The background color reflects a 60-minute rolling average SHAP value. (Bottom) The spearman correlation to show the association of the 60-minute rolling average actigraphy data and SHAP values mapped to a 24-hour cycle.*

4 Discussion

4.1 General Overview

Our work provides important contributions to the field of ambulatory mental health assessment and supports the use of a novel time-series analysis approach in this context. Our findings highlight the capacity of passively-collected, actigraphy data to model MDD in a large, nationally-representative sample. In particular, we find our self-organizing approach, which ingests passively-collected actigraphic time-series *agnostic to existing domain knowledge*, outperforms our traditional theory-organized approach, which leverages domain knowledge for the engineering of features related to movement, sleep and sedentary behavior. Additionally, we observed that the less-complex LSTM and Logistic Regression models performed only slightly better than chance. Further, introspection of the self-organizing model highlighted the importance of nighttime sleep disturbance for the model's predictions of depressed individuals; higher mid-morning and lower evening activity were influential for the model's detection of nondepressed individuals. Our study finds promising leads for future exploratory efforts aimed at expanding and refining measurement in ambulatory depression assessment.

4.2 Implications and Importance

By modeling MDD with both summative and minute-to-minute representations of the actigraphy data, we investigated physical activity phenotypes of MDD at varying levels of temporal resolution. As such, we observed that depressed individuals are less active overall, with the largest difference in movement intensity occurring in the early-to-late morning. This finding supports prior research indicating that depressed individuals are more sedentary and less active than nondepressed individuals (Gianfredi et al., 2022; Minaeva et al., 2020). The lower activity may be a proxy marker for other symptoms, as individuals endorsing sleep disturbances, fatigue, and anhedonia are less active (Leventhal, 2012). We observed distinct nighttime patterns of increased activity among the depressed group, consistent with prior research (Burton et al., 2013; Price et al., 2022; Rykov et al., 2021), perhaps indicating sleep disturbances (e.g., insomnia). Taken together, our findings suggest that depressed individuals have less daytime and more nighttime activity, and actigraphy may assist in indirectly capturing other depressive symptoms (e.g., sleep disturbances) beyond self-report measures.

Our work supports the use of actigraphy sensors coupled with deep learning to better understand the role of movement and sleep in MDD. In line with current literature (Gianfredi et al., 2022; Minaeva et al., 2020), our findings provide support that less overall movement is related to higher depressive symptoms and provide important clinical implications. Existing treatments, including behavioral activation, target activity levels and depressive symptoms by implementing activity scheduling to improve one’s mood (Cuijpers et al., 2007). Depressed individuals monitor their mood and activity to determine a link between the two (e.g., if certain activities lead to changes in mood). Several digital interventions have been developed with a behavioral activation framework, making this treatment more accessible (Huguet et al., 2016). Future research should investigate whether it would be beneficial for depressed individuals to utilize this intervention in the early morning, as we observed the largest differences in activity between the outcome groups during these hours.

Lastly, the implications of our work are important when considered in the context of just-in-time adaptive interventions (JITAIs), aimed at providing the right support at the right time, given one's current state (Hardeman et al., 2019; Nahum-Shani et al., 2015; L. Wang & Miller, 2020). The effective means by which to continuously and unobtrusively monitor an individual's emotional and context state is a prerequisite for the development of JITAIs. Consider, for instance, a JIT behavioral activation intervention aimed at improving depressive symptoms; to have optimal effect, such an intervention should be delivered to persons who are experiencing depressive symptoms. Our work provides early support for the use of unobtrusive sensor data to detect MDD using a screener questionnaire validated for MDD, which could then be acted upon by a tailored digital intervention.

4.3 Strengths and Limitations

The present study utilizes one of the largest, nationally-representative data sources, NHANES, which includes actigraphic and PHQ-9 data from individuals with wrist-worn accelerometers. Methodologically, our study provides a direct comparison of a theory-organized and self-organizing machine learning modeling framework to detect MDD via the PHQ-9, providing an in-depth analysis of the relationship between activity and depression.

However, our study has some important limitations. First, we utilized the PHQ-9 to indicate MDD, which may not have as strong validity as other forms of assessment, such as clinical interviews. Moreover, the PHQ-9 is a screening measure for MDD and not an assessment tool for formally diagnosing MDD. Thus, our findings are limited such that actigraphy could detect MDD *screening* rather than *diagnosis*. Future research should aim to investigate whether these findings replicate in detecting a diagnosis of MDD (e.g., using a clinical interview). Second, our study did not use imputation for actigraphy values predicted as non-valid or non-wear scores; however, it is possible that device compliance behaviors are influential in the model’s signal. Last, this study did not investigate other psychiatric outcomes, such as anxiety, which often co-occur with MDD, nor did the study investigate distinct depressive symptoms. Given that MDD is a highly comorbid and heterogeneous classification, future work should investigate the capacity of actigraphy to model comorbid symptoms, and individual depressive symptoms.

4.4 Conclusions

Our work demonstrates considerable promise in the use of passively-collected actigraphy data paired with machine and deep learning frameworks to detect MDD. Methodologically, our work provides a direct comparison between a domain-theory-integrated approach and a domain-agnostic, data-driven approach, and we find comparable, but superior performance of the data-driven approach. Our work uses the largest known available nationally-representative dataset, and contributes to ongoing research examining the utility of passive, unobtrusive data types in understanding depression.

Declarations of interest: none

References

Akincigil, A., & Matthews, E. B. (2017). National Rates and Patterns of Depression Screening in Primary Care: Results From 2012 and 2013. *Psychiatric Services*, *68*(7), 660–666. https://doi.org/10.1176/appi.ps.201600096

Ameko, M. K., Cai, L., Boukhechba, M., Daros, A., Chow, P. I., Teachman, B. A., Gerber, M. S., & Barnes, L. E. (2018). Cluster-based approach to improve affect recognition from passively sensed data. *2018 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*, 434–437. https://doi.org/10.1109/BHI.2018.8333461

Bailey, A. P., Hetrick, S. E., Rosenbaum, S., Purcell, R., & Parker, A. G. (2018). Treating depression with physical activity in adolescents and young adults: A systematic review and meta-analysis of randomised controlled trials. *Psychological Medicine*, *48*(7), 1068–1083. https://doi.org/10.1017/S0033291717002653

Belcher, B. R., Wolff-Hughes, D. L., Dooley, E. E., Staudenmayer, J., Berrigan, D., Eberhardt, M. S., & Troiano, R. P. (2021). US Population-referenced Percentiles for Wrist-Worn Accelerometer-derived Activity. *Medicine & Science in Sports & Exercise*, *53*(11), 2455–2464. https://doi.org/10.1249/MSS.0000000000002726

Biddle, D. J., Robillard, R., Hermens, D. F., Hickie, I. B., & Glozier, N. (2015). Accuracy of self-reported sleep parameters compared with actigraphy in young people with mental ill-health. *Sleep Health*, *1*(3), 214–220. https://doi.org/10.1016/j.sleh.2015.07.006

Burton, C., McKinstry, B., Szentagotai Tătar, A., Serrano-Blanco, A., Pagliari, C., & Wolters, M. (2013). Activity monitoring in patients with depression: A systematic review. *Journal of Affective Disorders*, *145*(1), 21–28. https://doi.org/10.1016/j.jad.2012.07.001

Camacho, T. C., Roberts, R. E., Lazarus, N. B., Kaplan, G. A., & Cohen, R. D. (1991). Physical

Activity and Depression: Evidence from the Alameda County Study. *American Journal of Epidemiology*, *134*(2), 220–231. https://doi.org/10.1093/oxfordjournals.aje.a116074

Case, A., & Deaton, A. (2015). Rising morbidity and mortality in midlife among white non-Hispanic Americans in the 21st century. *Proceedings of the National Academy of Sciences*, *112*(49), 15078–15083. https://doi.org/10.1073/pnas.1518393112

Centers for Disease Control and Prevention. (2005). *NHANES Analytic and Reporting Guidelines*. https://www.cdc.gov/nchs/data/nhanes/nhanes_03_04/nhanes_analytic_guidelines_dec_2005.pdf

Centers for Disease Control and Prevention. (2007). *NHANES 2003-2004: Physical Activity Monitor Data Documentation, Codebook, and Frequencies.* https://wwwn.cdc.gov/nchs/nhanes/2003-2004/PAXRAW_C.htm#PAXSTAT

Centers for Disease Control and Prevention. (2020a). *NHANES 2011-2012 Examination Data Overview*. Https://Wwwn.Cdc.Gov/Nchs/Nhanes/Search/Datapage.Aspx?Component=Examination&CycleBeginYear=2011.

Centers for Disease Control and Prevention. (2020b). *NHANES 2013-2014 Examination Data Overview*. Https://Wwwn.Cdc.Gov/Nchs/Nhanes/Continuousnhanes/Overviewexam.Aspx?BeginYear=2013.

Centers for Disease Control and Prevention. (2022). *Data Quality Flag Summary Table for the Physical Activity Monitor (PAM) Data Collected in NHANES 2011-2014 and NNYFS*. https://wwwn.cdc.gov/nchs/nhanes/Pam/Default.aspx

Cuijpers, P., van Straten, A., & Warmerdam, L. (2007). Behavioral activation treatments of depression: A meta-analysis. *Clinical Psychology Review*, *27*(3), 318–326. https://doi.org/10.1016/j.cpr.2006.11.001

DeMasi, O., & Recht, B. (2017). A step towards quantifying when an algorithm can and cannot predict an individual's wellbeing. *Proceedings of the 2017 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2017 ACM International Symposium on Wearable Computers*, 763–771. https://doi.org/10.1145/3123024.3125609

Franzen, P. L., & Buysse, D. J. (2008). Sleep disturbances and depression: Risk relationships for subsequent depression and therapeutic implications. *Dialogues in Clinical Neuroscience, 10*(4), 473–481. https://doi.org/10.31887/DCNS.2008.10.4/plfranzen

Gao, S., Calhoun, V. D., & Sui, J. (2018). Machine learning in major depression: From classification to treatment outcome prediction. *CNS Neuroscience & Therapeutics*, *24*(11), 1037–1052. https://doi.org/10.1111/cns.13048

Ghio, L., Gotelli, S., Cervetti, A., Respino, M., Natta, W., Marcenaro, M., Serafini, G., Vaggi, M., Amore, M., & Belvederi Murri, M. (2015). Duration of untreated depression influences clinical outcomes and disability. *Journal of Affective Disorders*, *175*, 224–228. https://doi.org/10.1016/j.jad.2015.01.014

Gianfredi, V., Schaper, N. C., Odone, A., Signorelli, C., Amerio, A., Eussen, S. J. P. M., Köhler, S., Savelberg, H. H. C. M., Stehouwer, C. D. A., Dagnelie, P. C., Henry, R. M. A., van der Kallen, C. J. H., van Greevenbroek, M. M. J., Schram, M. T., & Koster, A. (2022). Daily patterns of physical activity, sedentary behavior, and prevalent and incident depression—The Maastricht Study. *Scandinavian Journal of Medicine & Science in*

Sports, *32*(12), 1768–1780. https://doi.org/10.1111/sms.14235

Hallyburton, A., & Allison-Jones, L. (2023). Mental health bias in physical care: An integrative review of the literature. *Journal of Psychiatric and Mental Health Nursing*, jpm.12911. https://doi.org/10.1111/jpm.12911

Hardeman, W., Houghton, J., Lane, K., Jones, A., & Naughton, F. (2019). A systematic review of just-in-time adaptive interventions (JITAIs) to promote physical activity. *International Journal of Behavioral Nutrition and Physical Activity*, *16*(1), 31. https://doi.org/10.1186/s12966-019-0792-7

Heinz, M. V., Price, G. D., Ruan, F., Klein, R. J., Nemesure, M., Lopez, A., & Jacobson, N. C. (2022). Association of Selective Serotonin Reuptake Inhibitor Use With Abnormal Physical Movement Patterns as Detected Using a Piezoelectric Accelerometer and Deep Learning in a Nationally Representative Sample of Noninstitutionalized Persons in the US. *JAMA Network Open*, *5*(4), e225403. https://doi.org/10.1001/jamanetworkopen.2022.5403

Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Hudson, K. L., & Collins, F. S. (2017). The 21st Century Cures Act—A View from the NIH. *New England Journal of Medicine*, *376*(2), 111–113. https://doi.org/10.1056/NEJMp1615745

Huguet, A., Rao, S., McGrath, P. J., Wozney, L., Wheaton, M., Conrod, J., & Rozario, S. (2016). A Systematic Review of Cognitive Behavioral Therapy and Behavioral Activation Apps for Depression. *PLOS ONE*, *11*(5), e0154248. https://doi.org/10.1371/journal.pone.0154248

Hung, C.-I., Yu, N.-W., Wu, K.-Y., Yang, C.-H., & Liu, C.-Y. (2015). The impact of the duration of an untreated episode on improvement of depression and somatic symptoms. *Neuropsychiatric Disease and Treatment*, 2245. https://doi.org/10.2147/NDT.S89498

Janiesch, C., Zschech, P., & Heinrich, K. (2021). Machine learning and deep learning. *Electronic Markets*, *31*(3), 685–695. https://doi.org/10.1007/s12525-021-00475-2

John, D., Tang, Q., Albinali, F., & Intille, S. (2019). An Open-Source Monitor-Independent Movement Summary for Accelerometer Data Processing. *Journal for the Measurement of Physical Behaviour*, *2*(4), 268–281. https://doi.org/10.1123/jmpb.2018-0068

Johnson, J. M., & Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance. *Journal of Big Data*, *6*(1), 27. https://doi.org/10.1186/s40537-019-0192-5

Kennedy, S. H. (2022). Core symptoms of major depressive disorder: Relevance to diagnosis and treatment. *Dialogues in Clinical Neuroscience*, *10*(3), 271–277. https://doi.org/10.31887/DCNS.2008.10.3/shkennedy

Kessler, R. C., Berglund, P., Demler, O., Jin, R., Koretz, D., Merikangas, K. R., Rush, A. J., Walters, E. E., & Wang, P. S. (2003). The Epidemiology of Major Depressive Disorder: Results From the National Comorbidity Survey Replication (NCS-R). *JAMA*, *289*(23), 3095. https://doi.org/10.1001/jama.289.23.3095

Kim, J., Calhoun, V. D., Shim, E., & Lee, J.-H. (2016). Deep neural network with weight sparsity control and pre-training extracts hierarchical features and enhances classification performance: Evidence from whole-brain resting-state functional connectivity patterns of schizophrenia. *NeuroImage*, *124*, 127–146. https://doi.org/10.1016/j.neuroimage.2015.05.018

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, *60*(6), 84–90. https://doi.org/10.1145/3065386

Kroenke, K., Spitzer, R. L., & Williams, J. B. W. (2001a). The PHQ-9. *Journal of General Internal Medicine*, *16*(9), 606–613. https://doi.org/10.1046/j.1525-1497.2001.016009606.x

Kroenke, K., Spitzer, R. L., & Williams, J. B. W. (2001b). The PHQ-9: Validity of a brief depression severity measure. *Journal of General Internal Medicine*, *16*(9), 606–613. https://doi.org/10.1046/j.1525-1497.2001.016009606.x

Kushki, A., Anagnostou, E., Hammill, C., Duez, P., Brian, J., Iaboni, A., Schachar, R., Crosbie, J., Arnold, P., & Lerch, J. P. (2019). Examining overlap and homogeneity in ASD, ADHD, and OCD: A data-driven, diagnosis-agnostic approach. *Translational Psychiatry*, *9*(1), 318. https://doi.org/10.1038/s41398-019-0631-2

Lekkas, D., Price, G., McFadden, J., & Jacobson, N. C. (2021). The application of machine learning to online mindfulness intervention data: A primer and empirical example in compliance assessment. *Mindfulness*, *12*(10), 2519–2534. https://doi.org/10.1007/s12671-021-01723-4

Leventhal, A. M. (2012). Relations between anhedonia and physical activity. *American Journal of Health Behavior*, *36*(6), 860–872. https://doi.org/10.5993/AJHB.36.6.12

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 4765–4774.

Mammen, G., & Faulkner, G. (2013). Physical activity and the prevention of depression: A systematic review of prospective studies. *American Journal of Preventive Medicine*, *45*(5), 649–657. https://doi.org/10.1016/j.amepre.2013.08.001

Manea, L., Gilbody, S., & McMillan, D. (2015). A diagnostic meta-analysis of the Patient Health Questionnaire-9 (PHQ-9) algorithm scoring method as a screen for depression. *General Hospital Psychiatry*, *37*(1), 67–75. https://doi.org/10.1016/j.genhosppsych.2014.09.009

Marques, A., Bordado, J., Peralta, M., Gouveia, E. R., Tesler, R., Demetriou, Y., & Gomez Baya, D. (2020). Cross-sectional and prospective relationship between physical activity and depression symptoms. *Scientific Reports*, *10*(1), 16114. https://doi.org/10.1038/s41598-020-72987-4

Maurice-Tison, S., Verdoux, H., Gay, B., Perez, P., Salamon, R., & Bourgeois, M. L. (1998). How to improve recognition and diagnosis of depressive syndromes using international diagnostic criteria. *The British Journal of General Practice: The Journal of the Royal College of General Practitioners*, *48*(430), 1245–1246.

Meinlschmidt, G., Tegethoff, M., Belardi, A., Stalujanis, E., Oh, M., Jung, E. K., Kim, H.-C., Yoo, S.-S., & Lee, J.-H. (2020). Personalized prediction of smartphone-based psychotherapeutic micro-intervention success using machine learning. *Journal of Affective Disorders*, *264*, 430–437. https://doi.org/10.1016/j.jad.2019.11.071

Mellem, M. S., Liu, Y., Gonzalez, H., Kollada, M., Martin, W. J., & Ahammad, P. (2020). Machine Learning Models Identify Multimodal Measurements Highly Predictive of Transdiagnostic Symptom Severity for Mood, Anhedonia, and Anxiety. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *5*(1), 56–67. https://doi.org/10.1016/j.bpsc.2019.07.007

Minaeva, O., Booij, S. H., Lamers, F., Antypa, N., Schoevers, R. A., Wichers, M., & Riese, H. (2020). Level and timing of physical activity during normal daily life in depressed and non-depressed individuals. *Translational Psychiatry*, *10*(1), 259.

https://doi.org/10.1038/s41398-020-00952-w

Mojtabai, R., Olfson, M., & Han, B. (2016). National Trends in the Prevalence and Treatment of Depression in Adolescents and Young Adults. *PEDIATRICS*, *138*(6), e20161878–e20161878. https://doi.org/10.1542/peds.2016-1878

Moriarty, A. S., Gilbody, S., McMillan, D., & Manea, L. (2015). Screening and case finding for major depressive disorder using the Patient Health Questionnaire (PHQ-9): A meta-analysis. *General Hospital Psychiatry*, *37*(6), 567–576. https://doi.org/10.1016/j.genhosppsych.2015.06.012

Nahum-Shani, I., Hekler, E. B., & Spruijt-Metz, D. (2015). Building health behavior models to guide the development of just-in-time adaptive interventions: A pragmatic framework. *Health Psychology*, *34*(Suppl), 1209–1219. https://doi.org/10.1037/hea0000306

Nemesure, M. D., Heinz, M. V., Klein, R., McFadden, J., & Jacobson, N. C. (2021). *Predictive Modeling Approach to Evaluate Individual Response to a Physical Activity Digital Intervention for Subjects with Major Depressive Disorder*. PsyArXiv. https://doi.org/10.31234/osf.io/3kjyh

Nguyen, T., O'Dea, B., Larsen, M., Phung, D., Venkatesh, S., & Christensen, H. (2017). Using linguistic and topic analysis to classify sub-groups of online depression communities. *Multimedia Tools and Applications*, *76*(8), 10653–10676. https://doi.org/10.1007/s11042-015-3128-x

Onnela, J.-P., & Rauch, S. L. (2016). Harnessing Smartphone-Based Digital Phenotyping to Enhance Behavioral and Mental Health. *Neuropsychopharmacology*, *41*(7), 1691–1696. https://doi.org/10.1038/npp.2016.7

Opoku Asare, K., Terhorst, Y., Vega, J., Peltonen, E., Lagerspetz, E., & Ferreira, D. (2021).

Predicting Depression From Smartphone Behavioral Markers Using Machine Learning Methods, Hyperparameter Optimization, and Feature Importance Analysis: Exploratory Study. *JMIR mHealth and uHealth*, *9*(7), e26540. https://doi.org/10.2196/26540

Ozcift, A., & Gulten, A. (2011). Classifier ensemble construction with rotation forest to improve medical diagnosis performance of machine learning algorithms. *Computer Methods and Programs in Biomedicine*, *104*(3), 443–451. https://doi.org/10.1016/j.cmpb.2011.03.018

Park, E.-H., & Jung, M. H. (2019). The impact of major depressive disorder on adaptive function: A retrospective observational study. *Medicine*, *98*(52), e18515. https://doi.org/10.1097/MD.0000000000018515

Patterson, S. M., Krantz, D. S., Montgomery, L. C., Deuster, P. A., Hedges, S. M., & Nebel, L. E. (1993). Automated physical activity monitoring: Validation and comparison with physiological and self-report measures. *Psychophysiology*, *30*(3), 296–305. https://doi.org/10.1111/j.1469-8986.1993.tb03356.x

Pham, T. D. (2021). Time–frequency time–space LSTM for robust classification of physiological signals. *Scientific Reports*, *11*(1), Article 1. https://doi.org/10.1038/s41598-021-86432-7

Price, G. D., Heinz, M. V., Zhao, D., Nemesure, M., Ruan, F., & Jacobson, N. C. (2022). An unsupervised machine learning approach using passive movement data to understand depression and schizophrenia. *Journal of Affective Disorders*, *316*, 132–139. https://doi.org/10.1016/j.jad.2022.08.013

Reddy, M. S. (2010). Depression: The disorder and the burden. *Indian Journal of Psychological Medicine*, *32*(1), 1–2. https://doi.org/10.4103/0253-7176.70510

Riemann, D., Berger, M., & Voderholzer, U. (2001). Sleep and depression — results from psychobiological studies: An overview. *Biological Psychology*, *57*(1–3), 67–103.

https://doi.org/10.1016/S0301-0511(01)00090-4

Rykov, Y., Thach, T.-Q., Bojic, I., Christopoulos, G., & Car, J. (2021). Digital Biomarkers for Depression Screening With Wearable Devices: Cross-sectional Study With Machine Learning Modeling. *JMIR mHealth and uHealth*, *9*(10), e24872. https://doi.org/10.2196/24872

Schuch, F. B., & Stubbs, B. (2019). The Role of Exercise in Preventing and Treating Depression: *Current Sports Medicine Reports*, *18*(8), 299–304. https://doi.org/10.1249/JSR.0000000000000620

Shahriyari, L. (2019). Effect of normalization methods on the performance of supervised learning algorithms applied to HTSeq-FPKM-UQ data sets: 7SK RNA expression as a predictor of survival in patients with colon adenocarcinoma. *Briefings in Bioinformatics*, *20*(3), 985–994. https://doi.org/10.1093/bib/bbx153

Skotte, J., Korshøj, M., Kristiansen, J., Hanisch, C., & Holtermann, A. (2014). Detection of Physical Activity Types Using Triaxial Accelerometers. *Journal of Physical Activity and Health*, *11*(1), 76–84. https://doi.org/10.1123/jpah.2011-0347

Tao, X., Chi, O., Delaney, P. J., Li, L., & Huang, J. (2021). Detecting depression using an ensemble classifier based on Quality of Life scales. *Brain Informatics*, *8*(1), 2. https://doi.org/10.1186/s40708-021-00125-5

Trost, S. G., McIver, K. L., & Pate, R. R. (2005). Conducting accelerometer-based activity assessments in field-based research. *Medicine and Science in Sports and Exercise*, *37*(11 Suppl), S531-543. https://doi.org/10.1249/01.mss.0000185657.86065.98

Vallance, J. K., Winkler, E. A. H., Gardiner, P. A., Healy, G. N., Lynch, B. M., & Owen, N. (2011). Associations of objectively-assessed physical activity and sedentary time with depression: NHANES (2005–2006). *Preventive Medicine*, *53*(4–5), 284–288. https://doi.org/10.1016/j.ypmed.2011.07.013

Van Rossum, G., & Drake, F. L. (2009). *Python 3 Reference Manual*. CreateSpace.

Wang, L., & Miller, L. C. (2020). Just-in-the-Moment Adaptive Interventions (JITAI): A Meta-Analytical Review. *Health Communication*, *35*(12), 1531–1544. https://doi.org/10.1080/10410236.2019.1652388

Wang, S., Pathak, J., & Zhang, Y. (2019). Using Electronic Health Records and Machine Learning to Predict Postpartum Depression. *Studies in Health Technology and Informatics*, *264*, 888–892. https://doi.org/10.3233/SHTI190351

Wang, Z., & Oates, T. (2015). *Imaging Time-Series to Improve Classification and Imputation* (arXiv:1506.00327). arXiv. http://arxiv.org/abs/1506.00327

Zhai, L., Zhang, Y., & Zhang, D. (2015). Sedentary behaviour and the risk of depression: A meta-analysis. *British Journal of Sports Medicine*, *49*(11), 705–709. https://doi.org/10.1136/bjsports-2014-093613

Declarations of interest: none