An Automated Machine Learning Pipeline for Monitoring and Forecasting Mobile Health Data

A Technical Report submitted to the Department of Engineering Systems and Environment

Presented to the Faculty of the School of Engineering and Applied Science

University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree
Bachelor of Science, School of Engineering

Meredith Grehan

Spring, 2021.

Technical Project Team Members

Anna Bonaquist

Joseph Keogh

Owen Haines

Neil Singh

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

Asfaneh Doryab, Department of Engineering Systems and Environment

An Automated Machine Learning Pipeline for Monitoring and Forecasting Mobile Health Data

Anna Bonaquist*, Meredith Grehan*, Owen Haines*, Joseph Keogh*, Tahsin Mullick*, Neil Singh*,
Sam Shaaban†, Ana Radovic‡, and Afsaneh Doryab*
* University of Virginia, Charlottesville, VA
‡ University of Pittsburgh, Pittsburgh, PA
† NuRelm, Pittsburgh, PA
* {acb2qv, mhg8gq, ojh2et, jgk7uf, tum7q, ns9wr, ad4ks}@virginia.edu
‡ ana.radovic@chp.edu
† sam@nurelm.com

***Abstract*—Mobile sensing and analysis of data streams collected from personal devices such as smartphones and fitness trackers have become useful tools to help health professionals monitor and treat patients outside of clinics. Research in mobile health has largely focused on feasibility studies to detect or predict a health status. Despite the development of tools for collection and processing of mobile data streams, such approaches remain ad hoc and offline. This paper presents an automated machine learning pipeline for continuous collection, processing, and analysis of mobile health data. We test this pipeline in an application for monitoring and predicting adolescents' mental health. The paper presents system engineering considerations based on an exploratory machine learning analysis followed by the pipeline implementation.**

I. Introduction

Technology advances and widespread use of smartphones and fitness trackers have made these devices targets for mobile health applications. Passive sensing capabilities embedded in smart devices provide the opportunity to track and monitor behavioral cues related to health and wellbeing.

Mobile health approaches using passive sensing have focused on two main processes, namely representation or modeling of data. Representation refers to the reporting or recording of mental health data in applications, many of which are web or mobile based. Users typically interact with such applications to either record health states or receive treatment for illness (e.g., [1], [2]). Modeling is the act of utilizing data or features to produce insights on a user's mental health via passive sensing. Specifically, modeling approaches focus on the application of machine learning algorithms and their potential role in predicting health outcomes such as depression or anxiety (e.g., [3]–[5]).

However, there is a dearth of work that integrates both the modeling and representation paradigms in an end-to-end product through processing and modeling raw sensor data and communicating the insights to the stakeholders continuously. The focus of this work is on creating such a data pipeline, an automated suite, that will allow for raw mobile device data to predict health outcomes such as depression.

In the following section we describe the essential components of the pipeline and discuss existing tools and approaches to build each component. We then present an exploratory machine learning analysis that informs the implementation. We demonstrate the practicality of the pipeline with a use case on data from a group of adolescents facing depression.

This project is funded under NIMH project 1R44MH122067-01. *(Anna Bonaquist, Meredith Grehan, Owen Haines, Joseph Keogh, Tahsin Mullick, and Neil Singh contributed equally to the work.)*

II. Automated Processing and Prediction Pipeline

This section outlines the engineering design considerations for an automated processing and prediction pipeline for mobile health data. In the first part, we describe the architectural components of such a system including:

1) strategies for collection of subjective and objective behavioral data
2) methods for processing and preparing data for modeling
3) machine learning and analysis methods suited for modeling passive time series data
4) strategies for communicating results and insights back to the stakeholders (e.g., patients or clinicians)

In the second part, we demonstrate the feasibility of the automated pipeline through a case study using data from adolescents with depression.

A. Data Collection

1) Passive Sensing: Different mobile data collection frameworks have been developed over the past few years including AWARE Framework [6], Intellicare [7], MindLamp [8], RADAR [9], Sensus [10], CARP [11] and Beiwe [12]. These tools can use both active and passive sensing to collect social, behavioral, and cognitive data from users. The sensor data and derived behaviors from the user's phone and wearable trackers are utilized via a mobile application. Example data sources including GPS, accelerometer, call logs, screen status, wifi, and bluetooth allow inferences on movement patterns, physical activities, social communications, and phone usage. The data is then stored in a database or cloud storage where further analysis can be performed. In this work, we use the AWARE

Framework, an open-source platform with both Android and iOS clients that has been used in numerous studies to investigate health conditions e.g., [13]–[17]. AWARE Framework also integrates a study module to allow for monitoring of data streams which reduces the implementation efforts of researchers and mobile health developers.

2) Self-reports: Mobile health applications often employ regular subjective assessments from users in form of diaries or surveys. The frequency of self-reports depend on the application and can range from several times a day to once every few months. Many applications use Ecological Momentary Assessment (EMA) [18] to trigger short surveys to collect information from users in certain situations. There are also a number of different software options that allow for online self reporting including survey platforms such as Qualtrics and REDCap. Ideally, mobile applications should be able to objectively assess the state of individuals from sensor data alone without relying on the subjective assessments that are prone to diverge from reality. However, for the purpose of feasibility demonstration, our pipeline allows flexibility for collection of self-report data from different sources. In our implementation, we use data from REDCap, a platform implemented in alignment with most human-subjects research privacy standards, such as HIPAA.

B. Data Processing

Raw data from mobile devices is often noisy and incomplete and needs to be cleaned. Additionally, the raw data is rarely in a usable format for direct interpretation. As such, the pipeline should include mechanisms for cleaning and aggregating the data. Mobile feature engineering [19], [20] has been useful in drawing important insights from data. A few tools, including Digital Biomarker Discovery Pipeline (DBDP) [21], Health Outcomes through Positive Engagement and Self-Empowerment (HOPES) [22], and Reproducible Analysis Pipeline for Data Streams (RAPIDS) [19] have been developed to process mobile data streams. We implement RAPIDS in our pipeline because of its distinct advantages compared to other alternatives. RAPIDS offers an open source modular feature extraction platform originally tailored for data collected with AWARE and Fitness trackers. It also allows for flexible segmentation of time series and supports extraction of over 300 combined features related to movement, physical activities, social communications, phone usage, and sleep, most of which are listed in [20].

C. Modeling

1) Machine Learning: Machine learning models can be applied to behavioral features to predict a health outcome, e.g., [3], [23], [24]. Most machine learning methods are supervised and use patients' self-reported status or the clinical assessment to train algorithms and build models for future prediction of health status. A wide variety of machine learning algorithms can be applied to mobile data for health prediction. In their review of smartphone-based passive sensing applications, Cornet and Holden [23] stated that support vector machines, Bayes classifiers, decision trees, random forests, and linear regression were the most popular machine learning models used for predicting the status of a patient. Meta algorithms such as random forests and gradient boosting combine bagging and boosting to improve the machine learning results. Ghandeharioun et al. [3] demonstrated that such ensemble based methods generalized better on the test set than non-ensemble based methods. For this reason, we choose to try a variety of these algorithms including XGBoost, XGBoostRF, Random Forests, ExtraTrees, Gradient Boosting, Adaboost, Light GBM, and Catboost to determine the best fit.

In addition to learning algorithms, the pipeline should support continuous model building and refinement as new data is added to the streams. Current approaches in this domain aim to automatically produce test set predictions for a new dataset, e.g., AutoML [25], Auto-Weka [26], and Auto sklearn [27]. The training set and testing strategy, however, can have a major impact on the outcome and affect the generalizability of a machine learning model. Doryab et al. [24] discussed two kinds of validation approaches: individual patient models and unified patient models. Individual patient models use only the patient's data to predict their health state, while unified patient models learn from the entire sample of patients to make health predictions. In our implementation, we first run an exploratory analysis of the target dataset (as described in section III-C) to evaluate the training and testing methods and the corresponding machine learning outcomes. This analysis informs the optimal strategies to be implemented in the pipeline.

D. Communication and Visualization

Previous efforts in communicating health data have revealed the importance of evaluating the needs of multiple stakeholder groups and designing to fit those needs. For example, Abdullah et al. [28] uses patients' smartphones to both actively and passively track daily rhythms and to provide effective feedback that can help patients maintain a regular daily rhythm. It also feeds this clinically valuable information back to patients' physicians so that the physician knows how the patient is doing. In the development of Monarca, Frost et al [29] handled the disparate needs of stakeholder groups through the development of separate platforms for communication. A mobile application was developed for patient use, which allowed patients to view their mood level as predicted by the machine learning algorithm, as well as the factors contributing to their mood. A mobile application provides a convenient way for users to view the information that they value quickly, demonstrated in the MoodRhythm application [28]. In the development of Monarca, a web portal was created for use by care providers and researchers which displayed data of multiple patients. This web portal also allowed for care providers to see which patients were most at risk and possibly in need of immediate intervention. For our pipeline, we also designed a mobile application and web portal to inform clinicians and researchers about behavioral features selected by the machine learning algorithm as well as the actual and predicted health

outcome (see figure 2). We evaluated the wireframe designs for the mobile application through usability studies with the actual patients. The usability discussion is out of the scope of this paper and is reported elsewhere.

III. Pipeline Implementation for Real-time Assessment of Mental Health in Adolescents

This section details the pipeline components as described in the previous section. We present the implemented architecture (Fig. 1) and modeling informed by the exploratory analysis of a sample dataset from adolescents with depression. The specific components of the implemented pipeline are as follows:

1) **Data Collection**:
   - Passive raw sensor data collected on a daily basis via AWARE and Fitbit are stored in an InfluxDB database.
   - Self-reports PHQ-9 survey for assessing depression are collected once a week through REDCap.
2) **Data Processing**: AWARE data is cleaned and fed into the RAPIDS framework along with a configuration file that specifies the processing settings for feature extraction.
3) **Machine Learning**: Models of depression states are built from extracted features using PHQ-9 scores as ground truth for comparison.
4) **Web Portal and Mobile Client**: The machine learning results including predicted depression score and important features included in the model are communicated on the web portal and mobile clients.

The pipeline is designed to run once a week to align with the frequency of the REDCap surveys. We use Python libraries including pandas, scikit-learn, PyYAML, and SQLAlchemy for development and host the pipeline on an Amazon EC2 instance to promote automation and scalability.

A. Data Collection

AWARE data was originally stored in an InfluxDB database. InfluxDB is based on InfluxQL, a database dialect designed for time series data. Since new data is added to InfluxDB on an ongoing basis, performing further processing task on this database would slow down both data storage and query. As such, we created a secondary database for processing and added the data to this database on a weekly basis. We used a MySQL database in this implementation because of the RAPIDS incompatibility with the InfluxDB format. The duplication of data was done in three steps: copying the schema, copying the data, and verifying the data schema.

In the first step, the schema was copied from the InfluxDB database to the MySQL database. This was done by querying both the InfluxDB 'fields' and 'tags', similar to traditional SQL 'columns' and 'primary keys' respectfully, for all tables in the InfluxDB. The fields and tags were then used to create the respective columns in the MySQL database for each table. This was done to ensure the schema stayed in-tact during the data copying process.

In the second step, the data itself was inserted on a weekly rolling basis. To prevent the overload of the InfluxDB server, queries were sent on a regular basis via Python's 'influxdb' package. Queries were sent by table, by user, and by single day.

In the final step, the schema of the query result was verified. An artifact of InfluxDB is that if all the data for a column is null, the column will be excluded from the query result. To correct this, using the copied schema, we populated any missing columns in our query result with null values. Each query result was then written to the MySQL database via pandas SQLAlchemy and pandas libraries.

For the mental health monitoring of adolescents, the PHQ-9 (Patient Health Questionnaire) [30] survey was implemented on REDCap. PHQ-9 includes nine questions and the scores range from 1 (no depression) to 27 (severe depression). To assess the level of depression, the scores are categorized as minimal depression (1-4), mild depression (5-9), moderate depression (10-14), moderately severe depression (15-19), and severe depression (20-27). The collection of this data was facilitated by the UPMC Children's Hospital of Pittsburgh. As described later, we used both PHQ-9 scores and depression categories in our exploratory machine learning to assess the feasibility of classification vs. regression for our pipeline implementation.

B. Data Processing

The data processing is outlined in the following steps:

1) Data preparation: To prepare the raw mobile data for feature extraction, we first removed redundant columns to match table schemes appropriate for RAPIDS. The modified tables were stored in a separate MySQL database.

2) Create master participant file: We used information from our prepared AWARE MySQL data to generate a master participant file that consolidates information on all participants in the study. This file serves as a reference for RAPIDS. It contains date, participant unique identifier, and device (cell phone) unique identifier information [1].

3) Run RAPIDS to extract features from individual users: RAPIDS is designed to process an entire dataset in one run. Our implementation required ongoing processing and feature extraction for each individual user. As such, we adjusted the settings of RAPIDS to process data from individual users. Our script modifies a centralized configuration file based on data present for each user. To verify a given feature can be extracted, the script queries relevant data tables. The RAPIDS script is then executed to extract weekly features from the prepared AWARE data, and the extracted features for each user are sent to a database for storage and compiling.

4) Obtain PHQ-9 scores: The PHQ-9 scores were calculated and stored in a file on a regular basis. Ideally, the self-report measures are read through the API. However, an issue in matching the timestamp of the survey data prevented us from

[1]We used RAPIDS version 0.4.3. Further information on the current version of RAPIDS, its configuration, and the behavioral features available can be found at https://www.rapids.science/latest/

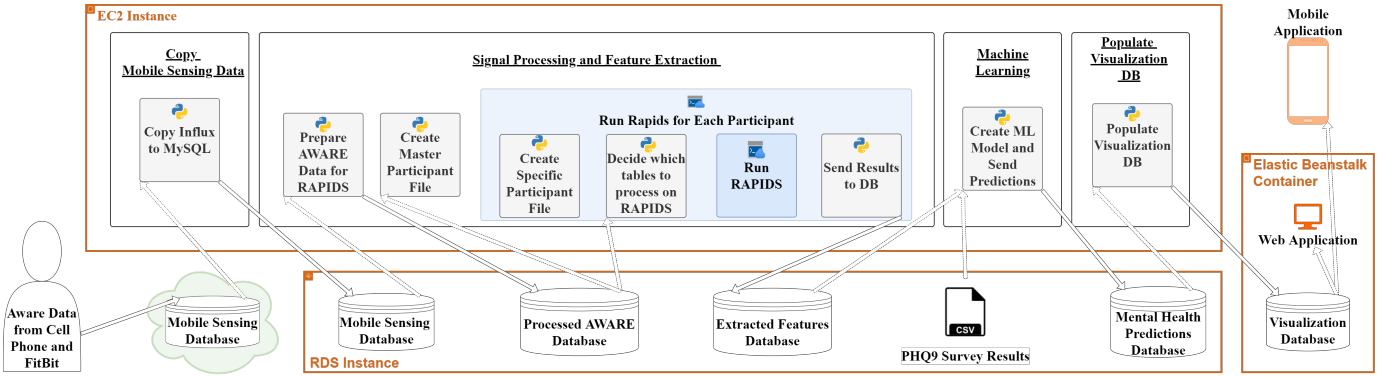

Fig. 1. Data Pipeline Implementation

using the REDCap API. Due to the lack of automation of this step, it has not been included in any of the figures. We matched the PHQ-9 scores and extracted features for each participants through a script on a weekly basis. Any extracted feature or survey result that did not have a match were dropped.

C. Machine Learning

With the data processing steps completed, the data was passed through a machine learning script to output depression predictions. RAPIDS outputs a large number of columns, many of which do not contain any data if patients turn off various sensors on their phones. In order to preserve useful information, all empty columns and rows, rows or columns with 50% or more missing data, and redundant columns with a correlation of over 0.95 were dropped to reduce multicollinearity. Median imputation was performed in the remaining dataset as many of the features were very skewed. The numerical columns were then scaled between 0 and 1 before training began.

To choose the final algorithms and validation strategy for implementation, we ran an exploratory analysis on data from 40 adolescents using different regression algorithms as listed in Table I using root mean square error (RMSE) as the performance measure.We chose XGBoostRF for the implementation as it provided the lowest RMSE. XGBoostRF stands for Extreme Gradient Boosting Random Forests and combines gradient boosting with an ensemble of decision trees to output a numerical prediction for depression score.

TABLE I
AVERAGE RMSE OF VARIOUS MACHINE LEARNING MODELS ON PATIENT DATA

Algorithm	Average RMSE
XGBoost	5.8
XGBoostRF	**5.3**
Random Forests	5.6
ExtraTrees	5.6
Gradient Boosting	5.8
Adaboost	5.4
Light GBM	6.0
Catboost	5.7

1) Validation Strategy and Performance Measures: Our approach used data aggregated on a weekly basis. Following the method in [24], we evaluated four validation strategies, two based on individual patient data and two based on the population data. The individual models include leave-one-patient-one-week-out (Lopowo) and leave-accumulated-weeks-out (Lawo). The former uses each patient's data alone to predict that patient's mental state for a given week. The latter uses a patient's data from weeks prior to the current week to predict the patient's mental state for that week. We also evaluated two unified patient model validation approaches namely leave-one-patient-out (Lopo) and leave-one-week-out (Lowo). Lopo uses all other patients' data to predict a given patient's mental state whereas Lowo uses all other weeks of data from all patients to predict a patient's mental state for that week.

To evaluate the automated modeling with continuous new data, we ran regression and classification experiments with the four validation strategies stated above. We first divided the data into nine sections each corresponding to one month of data. Each section added more recent data to the previous section, mimicking the increase in patient data collected by the app over time. The cross validation methods were tested across the nine increasing sections. Table II shows the results of each method using the XGBoostRF regressor and indicates that the Lawo strategy was the best validation approach for the data as it had the lowest RMSE on the final iteration. We therefore, chose to run the machine learning script on the accumulated patients' data on a weekly basis to train a model and to predict the depression score of the current week. The output of this model including the predicted depression score and the important features involved in modeling are stored in another database to be used on the web portal.

As previously mentioned, we also designed a mobile application for patients to monitor their mental health status. The design uses categories of depression (minimal to severe) instead of the actual PHQ-9 score. Therefore, we ran a classification test to choose the algorithm and the validation strategy. We initially converted the PHQ-9 scores into minimal, mild, moderate, moderately severe and severe.

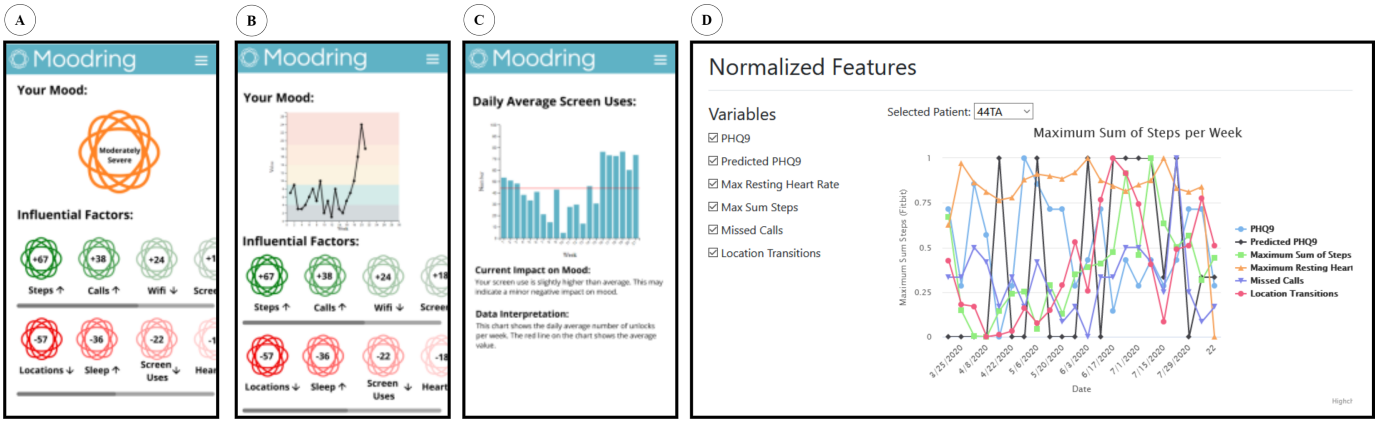

Fig. 2. Moodring App Preview

However, because of the sparsity in some categories and to balance the classes, we merged the categories to create labels "Minimal or Mild" (scores 0-9), "Moderate" (scores 10-14), and "Moderately Severe or Severe" (scores 15-27). We applied the classification version of XGBoostRF to predict the patient depression category. Each cross validation approach was tested using the classifier to see how they performed on iterations with increasing amounts of patient data (Table II). Consistent with the regression approach, the individual patient approaches show higher accuracy across all iterations than the unified patient validation approaches. This indicates that using personalized models for predicting patient PHQ-9 is more accurate than using models that pool data from a group of individuals.

TABLE II
XGBOOSTRF PREDICTION EVALUATION OVER INCREASING DATA

Iter.	Lopo		Lowo		Lopowo		Lawo	
	C Acc	R RMSE	C Acc	R RMSE	C Acc	R RMSE	C Acc	R RMSE
1	11%	3.8	50%	3.4	88%	2.7	86%	2.6
2	38%	4.6	70%	3.6	73%	3.1	77%	2.6
3	44%	5.6	70%	4.2	74%	3.5	74%	3.2
4	47%	6.3	67%	4.6	80%	3.5	80%	3.3
5	47%	5.9	64%	4.7	75%	3.5	75%	3.3
6	41%	5.7	62%	4.8	73%	3.3	73%	3.2
7	40%	5.7	60%	4.8	70%	3.4	71%	3.3
8	41%	5.7	59%	4.8	72%	3.5	73%	3.5
9	42%	5.8	59%	4.7	73%	3.4	73%	3.4

D. Visualization

The final step in the data pipeline was to automatically display and visualize the extracted features and machine learning predictions for further analysis by clinicians and researchers. A final product of the system would involve separate views and usability for members of the primary user classes of adolescents, parents, and care providers. Our current implementation is a dynamic web portal developed primarily for use as a developer and research portal seen in Fig. 2 D, but this implementation is also useful for clinicians to interpret the machine learning outputs. The front end is a Django-based application that is hosted on an Amazon Elastic Beanstalk system. The Django application accesses a database holding selected features and user PHQ-9 predictions and communicates these to the user by dynamically creating pages displaying line plots of normalized RAPIDS data for each feature used in the machine learning script separated by sensor. The features displayed are read from a generic configuration file that is updated based on the results of the machine learning The user has the ability to turn on and off each features in order to investigate specific trends as well as overlay the predicted and ground truth PHQ-9. In addition to the functional web portal, we also developed wireframes for a phone based application for adolescent users shown in Fig. 2 A, B and C. The design was refined through interviews with users and non-users. The mobile application utilizes feature importance from the machine learning algorithm and the classified PHQ-9 scores to display the predicted depression category as well behavioral features that contributed most to the prediction of that category. This gives users insights into what behavioral factors correspond to their depressive symptoms. The feature importance, direction, and intensity are presented using color, shade, and numerical labels.

IV. CONCLUSION AND FUTURE WORK

We presented an automated machine learning pipeline for continuous processing and analysis of mobile health data and communicating the results back to the stakeholders including clinicians, patients, and researchers. We implemented the pipeline for the real-time assessment of mental health in adolescents. Although we only used depression as a case for evaluation, the design and implementation can be extended to other health and behavioral outcomes with more biobehavioral data e.g., heart rate variability and skin temperature.

Future steps include allowing for simultaneous prediction of multiple outcomes, e.g., anxiety, stress, depression, and sleep quality. The pipeline can further be expanded to allow users to choose which combination of mental state forecasts they want to receive.

We plan to add functionality to allow clinicians and researchers to specify the analysis settings including selection

of the machine learning algorithm, identifying the prediction outcome, specifying the length of data to be used in the analysis, etc. This step will add more flexibility and value to the automated pipeline and help both clinicians and researchers draw insights in a ongoing basis.

We also plan on more in depth HCI research on the stakeholder interactions with the application to further identify key functionalities of the system. For example, an essential functionality for clinicians might be to give them an overview of patients at risk. This step will help the full implementation of the mobile application and web portal with added functionality and visualizations for all stakeholders.

REFERENCES

[1] R. F. Dickerson, E. I. Gorlin, and J. A. Stankovic, "Empath: A continuous remote emotional health monitoring system for depressive illness," in *Proceedings of the 2nd Conference on Wireless Health*, WH '11, (New York, NY, USA), Association for Computing Machinery, 2011.

[2] D. C. Mohr, K. N. Tomasino, E. G. Lattie, H. L. Palac, M. J. Kwasny, K. Weingardt, C. J. Karr, S. M. Kaiser, R. C. Rossom, L. R. Bardsley, L. Caccamo, C. Stiles-Shields, and S. M. Schueller, "Intellicare: An eclectic, skills-based app suite for the treatment of depression and anxiety," *J Med Internet Res*, vol. 19, p. e10, Jan 2017.

[3] A. Ghandeharioun, S. Fedor, L. Sangermano, D. Ionescu, J. Alpert, C. Dale, D. Sontag, and R. Picard, "Objective assessment of depressive symptoms with machine learning and wearable sensors data," in *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, pp. 325–332, 2017.

[4] A. Pratap, D. C. Atkins, B. N. Renn, M. J. Tanana, S. D. Mooney, J. A. Anguera, and P. A. Areán, "The accuracy of passive phone sensors in predicting daily mood," *Depression and Anxiety*, vol. 36, no. 1, pp. 72–81, 2019.

[5] A. A. Farhan, C. Yue, R. Morillo, S. Ware, J. Lu, J. Bi, J. Kamath, A. Russell, A. Bamis, and B. Wang, "Behavior vs. introspection: refining prediction of clinical depression via smartphone sensing data," in *2016 IEEE Wireless Health (WH)*, pp. 1–8, 2016.

[6] D. Ferreira, V. Kostakos, and A. K. Dey, "Aware: Mobile context instrumentation framework," *Frontiers in ICT*, vol. 2, p. 6, 2015.

[7] D. C. Mohr, K. N. Tomasino, E. G. Lattie, H. L. Palac, M. J. Kwasny, K. Weingardt, C. J. Karr, S. M. Kaiser, R. C. Rossom, L. R. Bardsley, L. Caccamo, C. Stiles-Shields, and S. M. Schueller, "Intellicare: An eclectic, skills-based app suite for the treatment of depression and anxiety," *J Med Internet Res*, vol. 19, p. e10, Jan 2017.

[8] J. Torous, H. Wisniewski, B. Bird, E. Carpenter, G. David, E. Elejalde, D. Fulford, S. Guimond, R. Hays, P. Henson, *et al.*, "Creating a digital health smartphone app and digital phenotyping platform for mental health and diverse healthcare needs: an interdisciplinary and collaborative approach," *Journal of Technology in Behavioral Science*, vol. 4, no. 2, pp. 73–85, 2019.

[9] Y. Ranjan, Z. Rashid, C. Stewart, P. Conde, M. Begale, D. Verbeeck, S. Boettcher, R. Dobson, and A. Folarin, "Radar-base: Open source mobile health platform for collecting, monitoring, and analyzing data using sensors, wearables, and mobile devices," *JMIR Mhealth Uhealth*, vol. 7, p. e11734, Aug 2019.

[10] H. Xiong, Y. Huang, L. E. Barnes, and M. S. Gerber, "Sensus: A cross-platform, general-purpose system for mobile crowdsensing in human-subject studies," in *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '16, (New York, NY, USA), p. 415–426, Association for Computing Machinery, 2016.

[11] J. E. Bardram, "The carp mobile sensing framework – a cross-platform, reactive, programming framework and runtime environment for digital phenotyping," 2020.

[12] J. Torous, M. V. Kiang, J. Lorme, and J.-P. Onnela, "New tools for new research in psychiatry: A scalable and customizable platform to empower data driven smartphone research," *JMIR Mental Health*, vol. 3, p. e16, May 2016.

[13] J. Vega, "Monitoring parkinson's disease progression using behavioural inferences, mobile devices and web technologies," in *Proceedings of the 25th International Conference Companion on World Wide Web*, WWW '16 Companion, (Republic and Canton of Geneva, CHE), p. 323–327, International World Wide Web Conferences Steering Committee, 2016.

[14] I. Moshe, Y. Terhorst, K. Opoku Asare, L. B. Sander, D. Ferreira, H. Baumeister, D. C. Mohr, and L. Pulkki-Råback, "Predicting symptoms of depression and anxiety using smartphone and wearable data," *Frontiers in Psychiatry*, vol. 12, p. 43, 2021.

[15] S. Bae, T. Chung, D. Ferreira, A. K. Dey, and B. Suffoletto, "Mobile phone sensors and supervised machine learning to identify alcohol use events in young adults: Implications for just-in-time adaptive interventions," *Addictive Behaviors*, vol. 83, pp. 42–47, 2018. Ambulatory Assessment of Addictive Disorders.

[16] A. Doryab, D. K. Villalba, P. Chikersal, J. M. Dutcher, M. Tumminia, X. Liu, S. Cohen, K. Creswell, J. Mankoff, J. D. Creswell, and A. K. Dey, "Identifying behavioral phenotypes of loneliness and social isolation with passive sensing: Statistical analysis, data mining and machine learning of smartphone and fitbit data," *JMIR Mhealth Uhealth*, vol. 7, p. e13209, Jul 2019.

[17] P. Chikersal, A. Doryab, M. Tumminia, D. K. Villalba, J. M. Dutcher, X. Liu, S. Cohen, K. G. Creswell, J. Mankoff, J. D. Creswell, M. Goel, and A. K. Dey, "Detecting depression and predicting its onset using longitudinal symptoms captured by passive sensing: A machine learning approach with robust feature selection," *ACM Trans. Comput.-Hum. Interact.*, vol. 28, Jan. 2021.

[18] S. Shiffman, A. A. Stone, and M. R. Hufford, "Ecological momentary assessment," *Annu. Rev. Clin. Psychol.*, vol. 4, pp. 1–32, 2008.

[19] J. Vega, M. Li, K. Aguillera, N. Goel, E. Joshi, K. C. Durica, A. R. Kunta, and C. A. Low, "Rapids: Reproducible analysis pipeline for data streams collected with mobile devices," 08 2020.

[20] A. Doryab, P. Chikarsel, X. Liu, and A. K. Dey, "Extraction of behavioral features from smartphone and wearable data," 2019.

[21] B. Brent, B. Lu, J. Kim, and J. P. Dunn, "Biosignal compression toolbox for digital biomarker discovery," *Sensors (14248220)*, vol. 21, no. 2, pp. 526–527, 2021.

[22] X. Wang, N. Vouk, C. Heaukulani, T. Buddhika, W. Martanto, J. Lee, and R. J. Morris, "Hopes: An integrative digital phenotyping platform for data collection, monitoring, and machine learning," *J Med Internet Res*, vol. 23, p. e23984, Mar 2021.

[23] V. P. Cornet and R. J. Holden, "Systematic review of smartphone-based passive sensing for health and wellbeing," *Journal of Biomedical Informatics*, vol. 77, pp. 120–132, 2018.

[24] A. Doryab, M. Frost, M. Faurholt-Jepsen, L. V. Kessing, and J. E. Bardram, "Impact factor analysis: combining prediction with parameter ranking to reveal the impact of behavior on health outcome," *Personal and Ubiquitous Computing*, vol. 19, no. 2, pp. 355–365, 2015.

[25] M. Feurer, A. Klein, K. Eggensperger, J. Springenberg, M. Blum, and F. Hutter, "Efficient and robust automated machine learning," in *Advances in Neural Information Processing Systems* (C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, eds.), vol. 28, Curran Associates, Inc., 2015.

[26] C. Thornton, F. Hutter, H. H. Hoos, and K. Leyton-Brown, "Auto-weka: Combined selection and hyperparameter optimization of classification algorithms," in *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, (New York, NY, USA), p. 847–855, Association for Computing Machinery, 2013.

[27] M. Feurer, A. Klein, K. Eggensperger, J. T. Springenberg, M. Blum, and F. Hutter, "Auto-sklearn: efficient and robust automated machine learning," *Automated Machine Learning*, pp. 113–134, 2018.

[28] S. Abdullah, M. Matthews, E. Frank, G. Doherty, G. Gay, and T. Choudhury, "Automatic detection of social rhythms in bipolar disorder," *Journal of the American Medical Informatics Association*, vol. 23, pp. 538–543, 03 2016.

[29] M. Frost, A. Doryab, M. Faurholt-Jepsen, L. V. Kessing, and J. E. Bardram, "Supporting disease insight through data analysis: refinements of the monarca self-assessment system," in *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*, pp. 133–142, 2013.

[30] 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.