A Computational Framework for Modeling Biobehavioral Rhythms from Mobile and Wearable Data Streams

RUNZE YAN, University of Virginia, USA
XINWEN LIU and JANINE DUTCHER, Carnegie Mellon University, USA
MICHAEL TUMMINIA, University of Pittsburgh, USA
DANIELLA VILLALBA, SHELDON COHEN, DAVID CRESWELL, and KASEY CRESWELL, Carnegie Mellon University, USA
JENNIFER MANKOFF and ANIND DEY, University of Washington, USA
AFSANEH DORYAB, University of Virginia, USA

This paper presents a computational framework for modeling biobehavioral rhythms - the repeating cycles of physiological, psychological, social, and environmental events - from mobile and wearable data streams. The framework incorporates four main components: mobile data processing, rhythm discovery, rhythm modeling, and machine learning. We evaluate the framework with two case studies using datasets of smartphone, Fitbit, and OURA smart ring to evaluate the framework's ability to (1) detect cyclic biobehavior, (2) model commonality and differences in rhythms of human participants in the sample datasets, and (3) predict their health and readiness status using models of biobehavioral rhythms. Our evaluation demonstrates the framework's ability to generate new knowledge and findings through rigorous micro- and macro-level modeling of human rhythms from mobile and wearable data streams collected in the wild and using them to assess and predict different life and health outcomes.

CCS Concepts: • **Human-centered computing → Ubiquitous and mobile computing**; • **Applied computing → Life and medical sciences;**

Additional Key Words and Phrases: Computational modeling, machine learning, biobehavioral rhythms, mental health, human behavior modeling

ACM Reference format:
Runze Yan, Xinwen Liu, Janine Dutcher, Michael Tumminia, Daniella Villalba, Sheldon Cohen, David Creswell, Kasey Creswell, Jennifer Mankoff, Anind Dey, and Afsaneh Doryab. 2022. A Computational Framework for Modeling Biobehavioral Rhythms from Mobile and Wearable Data Streams. *ACM Trans. Intell. Syst. Technol.* 13, 3, Article 47 (March 2022), 27 pages.
https://doi.org/10.1145/3510029

Authors' addresses: R. Yan and A. Doryab, University of Virginia, 1827 University Avenue, Charlottesville, Virginia, 22903, USA; emails: {ry4jr, ad4ks}@virginia.edu; X. Liu, J. Dutcher, D. Villalba, S. Cohen, D. Creswell, and K. Creswell, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, Pennsylvania, 15213, USA; emails: {xinwenl, jdutcher}@andrew.cmu.edu, daniella.villalba@gmail.com, {scohen, creswell}@cmu.edu, kasey@andrew.cmu.edu; M. Tumminia, University of Pittsburgh, 4200 Fifth Ave, Pittsburgh, Pennsylvania, 15260, USA; email: mjt94@pitt.edu; J. Mankoff and A. Dey, University of Washington, 1400 NE Campus Parkway, Seattle, Washington, 98195, USA; emails: {jmankoff, anind}@uw.edu.

2157-6904/2022/03-ART47
https://doi.org/10.1145/3510029

1 INTRODUCTION

The term biobehavioral rhythms introduced in [19], refers to the repeating cycles of physiological (e.g., heart rate and body temperature), psychological (e.g., mood), social (e.g., work events), and environmental (e.g., weather) that affect human body and life. Rooted in Chronobiology, "the scientific discipline that quantifies and explores the mechanisms of biological time structure and their relationship to the rhythmic manifestations in living matter" [15], biobehavioral rhythms aim at studying cyclic events observed in human data collected from personal and consumer level mobile and wearable devices [19]. Such devices provide the capability of continuous tracking of biobehavioral signals of individuals in their daily life and outside of controlled lab settings which have been the standard method for studying biological rhythms.

Numerous research studies have shown the impact of understanding rhythms and their effect on human life and wellbeing. For example, studies in [19, 28, 30] demonstrate the association between long-term disruption in biological rhythms and health outcomes such as cancer, diabetes, and depression. Other studies have shown the impact of shift work on the quality of life in shift workers such as nurses and doctors [33, 37]. These studies, however, have often been limited to controlled settings to observe certain behaviors and effects. With passive sensing of physiological and behavioral signals from mobile and wearable devices, it is now possible to study human rhythms more broadly and holistically in the wild through the collection of biobehavioral data from different sources. This opportunity, however, introduces new challenges. First, the longitudinal timeseries data collected from personal devices is massive, noisy, and incomplete requiring careful processing to extract and preserve useful fine-grained knowledge from data in various temporal granularity levels to be used for further modeling. Second, the fact that each data source (e.g., smartphone sensors) can capture different aspects of human rhythms (biological, behavioral, or both) requires exploration and incorporation of each signal to identify biological and behavioral indicators on the micro and macro level that may reveal a cyclic behavior. This process can be exhaustive and needs automation. Moreover, although the modeled rhythms by themselves can provide useful insights into human health and life, the exhaustive number of rhythm models generated by each source makes it difficult for manual interpretation of the models by researchers or experts. A further computational step should incorporate those models to provide further insights into different health and lifestyle outcomes both physical and mental.

We propose a computational framework to address the aforementioned challenges through a series of data processing and modeling steps. The framework first processes the raw sensor data collected from mobile and wearable devices to extract high-level features from those data streams. It then models biobehavioral rhythms for each sensor feature alone and in combination with other features to discover rhythmicity and other characteristics of cyclic behavior in the data. The biobehavioral rhythm models provide a series of characteristic features which are further used for measuring stability in biobehavioral rhythms and to predict different outcomes such as health status through a machine learning component. We evaluate the framework with two case studies. The first study uses mobile and Fitbit data collected from 138 college students over a semester to test the framework's ability to detect rhythmicity in students' data in different time frames over the course of the semester and to measure the stability and variation of rhythms among students with different mental health status. We then use the models of the rhythms to classify the mental health status of students at the end of the semester. The second study uses physio-behavioral data from 11 volunteers who wore OURA smart ring for 30 to 323 days. We test the framework's ability to detect long-term cycles in participants' biobehavioral data and to extract commonalities and differences in those cycles. We then use each person's significant cyclic periods in modeling individual rhythms and further predicting average daily readiness. Our research makes the following contributions:

(1) We introduce a computational framework for modeling biobehavioral rhythms to the mobile and ubiquitous computing community that provides the ability to a) flexibly process massive sensor data in different time granularity thus providing the ability to model and observe short- and long-term rhythmic behavior; b) identify variation and stability in individual and groups of time series data; and c) help observe the impact of cyclic biobehavioral parameters in revealing and predicting different outcomes (e.g., health).
(2) We demonstrate the framework's ability to generate new knowledge and findings via rigorous micro- and macro-level modeling of human rhythms from mobile and wearable data streams collected in the wild and using them to assess and predict different life and health outcomes.

In the following sections, we describe related work in the domain of mobile health and behavior modeling and discuss the motivation for modeling cyclic human behavior and its potential role in revealing health status. We then present our computational framework followed by case studies in modeling biobehavioral rhythms and exploring the role of those models in predicting mental health and readiness. We discuss the feasibility and flexibility of the framework in incorporating different analytic approaches and providing insights for building rhythm-aware technology.

2 BACKGROUND AND RELATED WORK

2.1 Biological Rhythms

The assessment of rhythmic phenomena in living organisms reveals the existence of events and behavior that repeat themselves in certain cycles and can be modeled with periodic functions [15, 54]. Each periodic function is specified by its average level, oscillation degree, and time of oscillation optimal. Biological rhythms, including patterns of activity and rest or circadian rhythms, have been extensively studied in Chronobiology and medicine [19, 28, 30] mostly in controlled environmental settings.

The advancements in activity trackers have made it possible to study these phenomena outside of the labs and have demonstrated the reliability of such devices in capturing circadian disruptions, including sleep and physical and mental health conditions. For example, studies using research grade actigraphy devices have shown differences in circadian rhythms among patients with bipolar disorder, ADHD, and schizophrenia [50]. Other studies have used the same type of data to explore circadian disruption in cancer patients undergoing chemotherapy [50]. Commercial devices such as Fitbits are now able to infer sleep duration and quality reasonably accurately. Two brief studies with healthy young adults have used activity data from Fitbit devices to quantify rest-activity rhythms and found that rhythm measurement compared well relative to research-grade actigraphy [5, 38]. Studies in [64] and [42] have also explored the capability of personal tracking devices to measure sleep compared to gold standards such as polysomnography.

2.2 Behavior Modeling in the Wild via Mobile Sensing

The study of biobehavioral rhythms also relates to research in understanding human behavior from passive sensing data collected via smartphones and wearable devices. Only few studies have actually used mobile data for understanding the circadian behavior of different chronotypes (e.g., [1–3]). Abdullah et al. [1] analyzed patterns of phone usage to demonstrate differences in the sleep behavior of early and late chronotypes. In a similar study using the same type of data, they showed the capability of using mobile data to explore daily cognition and alertness [2, 3] and found that body clock, sleep duration, and coffee intake impact alertness cycles.

Data from smartphones and wearable devices has extensively been used for modeling daily behavior patterns such as movement [17], sleep [45], and physical and social activities [47] to understand their associations with health and wellbeing. For example, Medan et al. [41] found that decreases in call, SMS messaging, Bluetooth-detected contacts, and location entropy (a measure of the popularity of various places) were associated with greater depression. Wang et al. [63] monitored 48 students' behavior data for one semester and demonstrated significant correlations between data from smartphones and students' mental health and educational performance. In addition, Saeb et al. [56] extracted features from GPS location and phone usage data and applied a correlation analysis to capture relationships between features and level of depression. They find that circadian movement (regularity of the 24h cycle of GPS change), normalized entropy (mobility between favorite locations), location variance (GPS mobility independent of location), phone usage features, usage duration, and usage frequency, were highly correlated with the depression score. Doryab et al. [20] studied loneliness detection through data mining and machine learning modeling of students' behavior from smartphone and Fitbit data and showed different patterns of behavior related to loneliness, including less time spent off-campus and in different academic facilities as well as less socialization during evening hours on weekdays among students with the high level of loneliness.

Recent tools such as Rhythomic [29] and ARGUS [31] use visualization to analyze human behavior. Rhythomic is an open-source R framework tool for general modeling of human behavior, including circadian rhythms. ARGUS, on the other hand, focuses on visual modeling of deviations in circadian rhythms and measures their degree of irregularity. Through multiple visualization panes, the tool facilitates the understanding of behavioral rhythms. This work is related to our computational framework for modeling human rhythms. However, in addition to the underlying assumption of, and a focus on, circadian rhythms only, these tools primarily enable understanding of rhythms through visualization, whereas in our framework, we provide means for processing different data sources, extracting information from them, and discovering and modeling rhythms for each biobehavioral signal with different periods other than 24 hours. To our knowledge, this is the first computational framework to extract and incorporate the parameters obtained from rhythm models in a machine learning pipeline to predict different outcomes.

3 COMPUTATIONAL FRAMEWORK FOR MODELING BIOBEHAVIORAL RHYTHMS

Our proposed framework (Figure 1) incorporates data streams from mobile and wearable devices, including behavioral signals such as movement, audio, Bluetooth, WiFi, and GPS and logs of phone usage and communication (calls and messages); and biosignals such as heart rate, skin temperature, and galvanic skin response. These signals are processed, and granular features that characterize biobehavioral patterns such as activity, sleep, social communication, work, and movements are extracted. The data streams of biobehavioral sensor features are segmented into different time windows of interest and sent to a rhythm discovery component that applies periodic functions on each windowed stream of the sensor feature to detect their periodicity. The detected periods are then used to model the rhythmic function that represents the time series data stream for that sensor feature. The parameters generated by the rhythmic function are used in two ways. First, they are aggregated and further processed to characterize the stability or variation in rhythms over a certain time segment. Second, they are used as features in a machine learning pipeline to predict an outcome of interest (e.g., health status). The following sections provide details on the methods used in different components of the framework.

3.1 Time Series Segmentation

Windowing is one of the most frequently used processing methods for streams of data. A time series of length L is split into N segments based on certain criteria such as time. Our framework allows different ways to segment the time series, including the widely used tumbling windows,

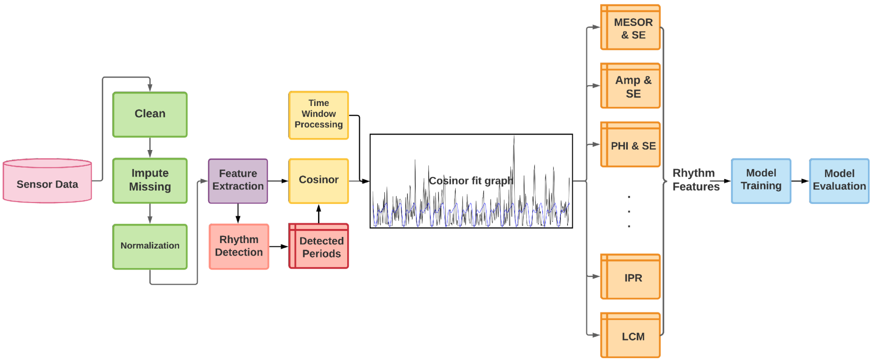


Fig. 1. Computational framework for modeling rhythms from mobile and wearable data streams and using the rhythm parameters for prediction of an outcome (e.g., health).

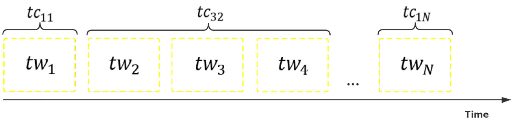


Fig. 2. The segmentation of time series with time windows (tw) and time chunks (tc).

which are a series of fixed-sized, non-overlapping and contiguous time intervals. We call each segment a time window (tw) which is a time series of length l, where $l = L/N$.

We also add a second segmentation layer to the time series where at each round k and starting point s ($s = 1 \ldots N$), we allow to combine a sequence of k consecutive time windows ($k = 1 \ldots N$) starting from time window s (tw_s) to generate time series of length k. We call these segments time chunks (tc). For example, in round $k = 1$, the tc_{11} is a time chunk of length one and starting point of tw_1 and tc_{12} is a time chunk of length one and starting point tw_2, whereas for $k = 3$, the tc_{32} is a time chunk of length three and starting point of tw_2. Time chunks allow flexible modeling of rhythms in different time periods over the length of the time series. Figure 2 illustrates the time segmentation process.

3.2 Detection of Rhythmicity

One of the first steps in modeling biobehavioral rhythms is identifying rhythmicity in time series data. We use two main methods for detecting and observing cyclic behavior: Autocorrelation and Periodogram.

3.2.1 Autocorrelation. Autocorrelation is a reliable analytical method for recognizing periodicities [21]. It calculates the correlation coefficient between a time series and its lagged version to measure their similarity over consecutive time intervals. Formally, the **autocorrelation function (ACF)** between two values y_t, y_{t-k} in a time series y_t is defined as

$$Corr(y_t, y_{t-k}), k = 1, 2, \ldots, \tag{1}$$

where k is the time gap and is called the lag [46]. In each iteration, the two time series are shifted by k points until one third of the data is parsed. If the time series is rhythmic, the coefficient values increase and decrease in regular intervals, and significant correlations indicate strong periodicity in data. The autocorrelation sequence of a periodic signal has the same cyclic characteristics as the signal itself. Thus, autocorrelation can help verify the presence of cycles and determine the

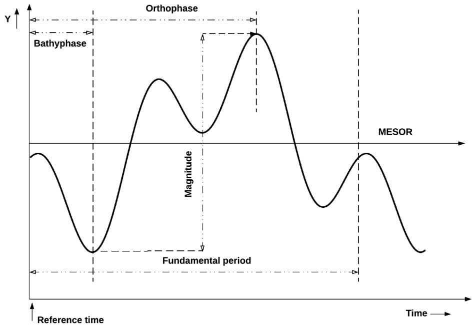


Fig. 3. Visualization of rhythm parameters [13].

periods. It has been empirically applied on various types of time series data from different fields and was shown to be dependable and exact in the tested situations [48, 57].

3.2.2 Periodogram. A key step in the rhythm discovery process is the estimation of the length of the period for each rhythm. Many different techniques and algorithms for determining the period of a cycle have been developed, including the Fourier-transform based methods such as Fast Fourier Transform [6], Non-Linear Least Squares [60], and Spectrum Resampling [14]. Other frequently used methods are Enright and Lomb-Scargle periodograms [24, 40], mFourfit [23], Maximum Entropy Spectral Analysis [11], and Chi-Square periodograms [59]. All of these methods come with different assumptions and with different levels of complexity [53]. For example, Spectrum Sampling has outperformed the usual Fourier approximation methods and has shown more robustness towards non-sinusoidal and noisy cycles [66]. It has also been used to detect changes in period length, which allows for the estimation of variance in different periods, as frequently observed in practice. These functionalities, however, have made the algorithm slow and computationally expensive [66].

Arthur Schuster used Fourier analysis to evaluate periodicity in meteorological phenomena and introduced the term *'periodogram'* [58]. The method was first applied to the study of circadian rhythms in the early 1950s to quantify free-running rhythms of mice after blinding [35]. Periodograms provide a measure of strength and regularity of the underlying rhythm through the estimation of the spectral density of a signal. For a time series $y_t, t = 1, 2, \ldots, T$, the spectral energy P_k of frequency k can be calculated as [52]:

$$P_k = \left(\frac{2}{T}\sum_{t=1}^{T} y_t cos\left(\frac{2\pi kt}{T}\right)\right)^2 + \left(\frac{2}{T}\sum_{t=1}^{T} y_t sin\left(\frac{2\pi kt}{T}\right)\right)^2 \tag{2}$$

The periodogram uses a Fourier Transform to convert a signal from the time domain to the frequency domain. A Fourier analysis is a method for expressing a function as a sum of periodic components and recovering the time series from those components. The dominant frequency corresponds to the periodicity in the pattern.

3.3 Modeling Rhythms

The next step in our framework is modeling the rhythmic behavior of a time series data, which is done via a periodic function. Each periodic function is among others specified by its period, average level (MESOR), oscillation degree (Amplitude), and time of oscillation optimal (Phase) [34]. The following rhythm parameters can be extracted from the model generated by the periodic function (Figure 3) [13, 25, 38]:

- *Fundamental period*: Periodic sequences are usually made up of multiple periodic components. The fundamental period measures the time during an overall cycle.
- *MESOR* is the midline of the oscillatory function. When the sampling interval is equal, the MESOR is equal to the mean value of all cyclic data points.
- *Amplitude (Amp)* refers to the maximum value a single periodic component can reach. The amplitude of a symmetrical wave is half of its range of up and down oscillation.
- *Magnitude* refers to the difference between the maximum value and the minimum value within a fundamental period. If a periodic sequence only contains one periodic component, amplitude equals half of the magnitude.
- *Acrophase (PHI)* refers to the time distance between the defined reference time point and the first time point in a cycle where the peak occurs with a period of a single periodic component.
- *Orthophase* refers to the time distance between the defined reference time point and the first time point in a cycle where the peak occurs with a fundamental period. When the time sequence only contains one periodic component, orthophase equals to acrophase.
- *Bathyphase* refers to the time distance between the defined reference time point and the first time point in a cycle where the trough occurs with a fundamental period.
- *P-value (P)* indicates the overall significance of the model fitted by a single period and comes from the F-test comparing the built model with the zero-amplitude model.
- *Percent rhythm (PR)* is the equivalent to the coefficient of determination (denoted by R^2) representing the proportion of overall variance accounted for by the fitted model.
- *Integrated p-value (IP)* represents the significance of the model fitted by the entire periods.
- *Integrated percent rhythm (IPR)* is the R^2 of the model fitted by the entire periods.
- *The longest cycle of the model (LCM)* equals to the least common multiple of all single periods.

The most fundamental method for modeling rhythms with known periods is Cosinor, a periodic regression function first developed by Halberg et al. [32] that uses the least-squares method to fit one or several cosine curves with or without polynomial terms to a single time series. It uses the following cosine function to model the time series [25]:

$$y_i = M + \sum_{c=1}^{C} A_c cos(\omega_c t_i + \phi_c) + e_i, \tag{3}$$

where y_i is the observed value at time t_i; M presents the MESOR; t_i is the sampling time; C is the set of all periodic components; A_c, ω_c, ϕ_c respectively presents the amplitude, frequency, and acrophase of each periodic component; and e_i is the error term. In addition to the parameters described above, Cosinor outputs the **standard error (SE)** for MESOR, amplitude, and acrophase, respectively.

The Cosinor models can be generated for one time series (single Cosinor - individual model) or for a group of time series (population-mean Cosinor - population model) through the aggregation of rhythm parameters obtained from single Cosinor. Cosinor models have been used to characterize circadian rhythms and compute relevant parameters with confidence limits. The model outputs the significance of the period, and it is proved that if $P \leq 0.05$, the assumed period actually exists. Our Cosinor framework allows for different periodic functions to be applied to the time series data using the detected periods from the previous step. We then use the rhythmic parameters measured by the Cosinor model in our machine learning pipeline as described in the next section.

3.4 Machine Learning Method

The machine learning component of the framework uses the parameters obtained from modeling the rhythm of each sensor feature to generate datasets for training and testing of an outcome of

ALGORITHM 1: Missing value imputation

```
Data: Input dataset $D$
Find the indexes list of the existing values $In$
Missing value counter: $c = In[0]$
for $i = 1$ to $len(In)$ do
    $index\_diff = In[i] - In[i-1]$
    if $index\_diff > 1$ then
        $value\_diff = D[In[i]] - D[In[i-1]]$
        $c = c + index\_diff$
        for $In[i-1] < j < In[i]$ do
            $D[j] = \frac{value\_diff}{index\_diff} \cdot (j - In[i])$
        end
    end
end
Missing rate threshold = $\theta$
Number of data points in $D = N$
if $\frac{c}{N} > \theta$ then
    Delete $D$
else
    return the imputed dataset
end
```

interest, e.g., health. The pipeline processes and handles missing values both in sensor and rhythm features across different time windows, selects important rhythm features as part of the training process, and builds machine learning models for the prediction of the outcome. The following sections describe the details of each step.

3.4.1 Handling Missing Values. Given the streams of data from multiple sources, the framework handles missing data for each sensor stream and each time window. We remove any sensor features if the percent of its missing data is greater than a threshold (e.g., 30%). For the remaining sensor features, we perform nearest-neighbor linear interpolation [8] to fill in missing values. For example, if there are three missing data points between 10 and 50, then those three missing points are filled with 20, 30, and 40, respectively. Given that the first and last data points cannot be imputed using this method, we remove the sensor feature if the first or the last data point in the time window is missing.

We apply the same process for handling missing rhythmic features in consecutive time windows. For each rhythmic feature, we fill the value of the missing time window with nearest-neighbor linear interpolation. Let v_i be the value of feature in time window tw_i. If v_1 and v_5, the values of features in time windows tw_1 and $tw_{5,}$ are present and v_2, v_3, and v_4, the feature values of tw_2, tw_3 and tw_4 are missing, then $diff = \frac{v_5 - v_1}{5-1}$, and $v_2 = v_1 + diff$, $v_3 = v_1 + diff * 2$, and $v_4 = v_1 + diff * 3$. For each missing time window, if none of the time windows before it has value, or none of the time windows after it has value, then this time window is not filled. After imputation, we remove any rhythmic feature with missing values more than a threshold (e.g., 30%). Algorithm 1 describes the process in more detail.

3.4.2 Feature Selection. As mentioned in previous sections, for each type of sensor feature, a single period or a multi-frequency Cosinor model is generated which outputs a list of rhythm parameters. These parameters are entered into the training process for building machine learning models.

Let M be the number of sensors ($s_1 \ldots s_m$), FN_i be the number of features for sensor i and RN_j the corresponding number of rhythmic features for feature j in sensor i. The resulting feature space will be of $M * FN * RN$ which is high dimensional compared to the relatively few data samples for training. As such, a reduction in the number of features is prevalent. The framework allows for the integration of different feature selection methods such as Lasso, **Randomized Logistic Regression (RLR)**, and **Information Gain (IG)** in the machine learning component.

Lasso is a linear regression model penalized with the L1 norm to fit the coefficients [10]. The Lasso regression prefers solutions with fewer non-zero coefficients and effectively reduces the number of features independent of the target variable. Through cross-validation, the lasso regression can output the importance level for each feature in the training dataset. We use a threshold value of 1e-5 to select features with Lasso, which is the default threshold in the scikit-learn library of Python [49]. Features with importance greater or equal to the threshold are kept, and the rest are discarded.

Randomized Logistic Regression is developed for stability selection of features. The basic idea behind stability selection is to use a base feature selection algorithm like logistic regression to find out which features are important in bootstrap samples of the original dataset [43]. The results on each bootstrap sample are then aggregated to compute a stability score for each feature in the data. Finally, features with a higher stability score than a threshold are selected. We use 0.25, the default threshold value in the scikit-learn library [49].

Information Gain (also referred to as Mutual Information in feature selection) measures the dependence between the features and the dependent variable (predicted outcome) [36]. Mutual information is always larger than or equal to zero, where the larger the value, the greater the relationship between the two variables. If the calculated result is zero, then the variables are independent. We set our algorithm to select 10 (the default value in the scikit-learn library [49]) features with highest information gain.

3.4.3 Model Building and Validation. The step for building machine learning models using rhythm features of k consecutive time windows and for a population of D data samples is flexible in the framework and can incorporate different supervised and unsupervised machine learning methods such as regression, classification, and clustering. In the current version of the framework, we implement three classification methods, including ***Logistic Regression (LR)***, ***Random Forest (RF)***, *and* ***Gradient Boosting (GB)***. The choice of algorithms is simply based on our empirical evidence of their performance on this type of data. Logistic regression [44] uses the logistic function to build a classifier. Random forest and Gradient Boosting are two branches of ensemble learning [16] which use the idea of bagging and boosting [9], respectively. Their common feature is to use the decision tree as the basic classifier and to get a robust model by combining multiple weak models. Bagging is short for boost strapped aggregation. Boost strapping is a repeated sampling method with replacement and random sampling [27]. In boosting, the training set of each iteration is unchanged, but the weight of samples is changed. At each iteration, the training samples with high error rates are given higher weights, so they get more attention in the next round of training.

We built two types of machine learning models: single sensor modeling and multiple sensor modeling. The single sensor model was built with rhythmic features extracted from a single sensor feature alone to better understand the contribution of each sensor feature in prediction. The multiple sensor model on the other hand was used to evaluate the combined power of multiple sensor features. We used a baseline of the majority class to measure the classifiers' performance in predicting the outcome. Again, the flexibility of the framework allows for the incorporation of different baseline measures. Both feature selection process and building machine learning models are done within a cross-validation setting, e.g., leave one sample out [65]. The machine learning component can measure basic performance measures of accuracy, precision, recall, F1, and MCC scores to evaluate the algorithms' performance. From those measures, we choose the results above baseline for each combination of feature selection and learning algorithm to further explore the prediction outcomes.

4 EVALUATION

To demonstrate the capability of our framework in building rhythm models from micro- and macro-level sensor features and utilizing them in prediction tasks, we present two different cases. The first case, utilizes data from smartphones and Fitbit to explore the relationship between biobehavioral rhythms and mental health status. The second case investigates long-term biobehavioral rhythms of data from OURA smart ring and their ability to predict readiness. We choose different analysis approaches to showcase the flexibility of the framework in handling different types of data and measuring various outcomes.

4.1 Case 1: Classification of Mental Health via Rhythm Models Using Data from Smartphone and Fitbit

We utilized a dataset of smartphones, Fitbit, and survey data collected from 138 first-year undergraduate students at an American university who were recruited for a health and well-being research study. The dataset was previously used in [20] to detect loneliness among college students. Smartphone data was collected through the AWARE framework [26] and included calls, messages, screen usage, Bluetooth, Wi-Fi, audio, and location. In addition, a Fitbit Flex2 wearable fitness tracker tracked steps, distances, calories burned, and sleep; and survey questions gathered information about physical and mental health including loneliness and depression. The survey data was collected at the beginning and at the end of the semester.

Our analysis was performed in two steps: First, we explored the potential of modeling and detecting rhythmicity in passively collected data from students' mobile and wearable data streams. Then, we used the built rhythm models to extract features that were fed into machine learning models to explore the relationship between students' biobehavioral rhythms and their mental health. We aimed to answer the following questions:

(1) Can we observe rhythmicity in students' biobehavioral data over the course of the semester? If so, are those rhythms consistent throughout the semester or do they change during different periods?
(2) Do we observe any difference in biobehavioral rhythms among students with different health status? If so, do healthy students have more stable rhythms?
(3) How accurately can models of biobehavioral rhythms predict mental health status?
(4) What are the most important characteristics and rhythmic features that reveal change in health status?

Note that our framework provides the ability to generate a large number of observations on the micro- (sensor feature) and macro-level (sensor), but in this paper, we only focus on observations related to our analysis questions.

4.1.1 Sensor Data Processing. The dataset collected from smartphones and Fitbits consisted of time series data from multiple sensors, including Bluetooth, calls, SMS, Wi-Fi, location, phone usage, steps, and sleep. We grouped this time series data into hourly bins and processed it following the approach in [18] to extract features related to mobility and activity patterns, communication and social interaction, and sleep. Examples of such features include travel distance, sleep efficiency, and movement intensity. We then split the semester data into tumbling cyclic time windows of

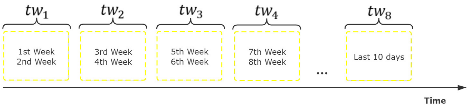


Fig. 4. The size of a time window is 2 weeks which segments the semester into roughly 8 time windows.

14 days or two weeks based on empirical evaluation of different lengths of time windows. The university semester in the studied population was roughly 16 weeks long, which could be divided into eight time windows of two weeks, except for the last time window that contained only ten days of data (Figure 4). We built a model of rhythm for each student and for each time window.

We handled missing sensor data on a per-participant per-time window basis. For each participant and each time window, we removed sensor features with more than 30% missing data. For the remaining sensor features, we performed nearest-neighbor linear interpolation as described previously to fill in missing values.

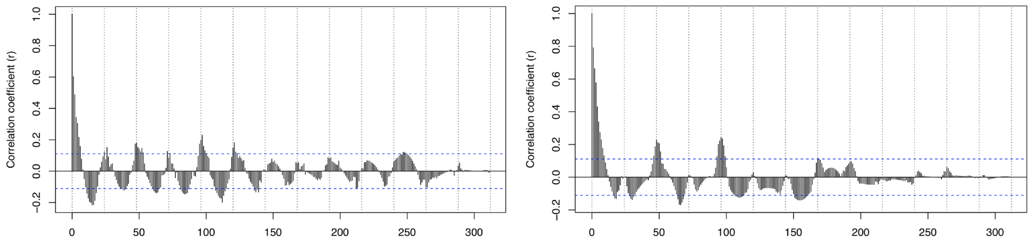
4.1.2 Ground Truth Measures for Loneliness and Depression. In our evaluation, we focused on two mental health outcomes, namely depression and loneliness. These two measures were chosen because of their longitudinal aspect, i.e., lasting for at least a few weeks to enable the investigation of 1) how biobehavioral rhythms of students with mental health conditions would differ from other students, and 2) how accurately the state of those mental health conditions could be predicted from extracted rhythms.

Loneliness data was collected using the UCLA Loneliness Scale, a well-validated and commonly used measure of general feelings of loneliness [55]. The questionnaire contains 20 questions about feeling lonely and isolated using a scale of 1 (never) to 4 (always). The total loneliness scores range from 20 to 80, with higher scores indicating higher levels of loneliness. As there is no standard cut-off for loneliness scores in the literature, we followed the same approach in [20] to divide the UCLA scores into two categories where the scores of 40 and below were categorized as *'low loneliness'*, and the scores above 40 were categorized as *'high loneliness'*.

Depression was assessed using the **Beck Depression Inventory-II (BDI-II)** [4, 22], a widely used psychometric test for measuring the severity of depressive symptoms that have been validated for college students [22]. The BDI-II contains 21 questions, with each answer being scored on a scale of 0-3 where higher scores indicate more severe depressive symptoms. For college students, the cut-offs on this scale are 0-13 (no or minimal depression), 14-19 (mild depression), 20-28 (moderate depression), and 29-63 (severe depression) [22]. For simplicity and to be consistent with the loneliness categorization, we divided these scores into two categories where the BDI-II scores <14 were labeled as *'not having depression'* and all BDI-II scores >= 14 were labeled as *'having depression'*.

Our machine learning pipeline used these loneliness and depression categories as ground truth labels to classify students' depression and loneliness levels using rhythmic features. Each student filled out the surveys both at the beginning (Pre) and the end of the semester (Post). To capture relationships between biobehavioral rhythms and changes in students' mental health, we categorized students into five groups according to the survey measures for depression and loneliness. For simplicity of representation, we further label *low loneliness* and *no depression* categories as 1, and *high loneliness* and *high depression* as 2. The five mental health categories are as follows:

- All students
- Pre1_ Post1: not having a mental health condition in both pre-semester and post-semester surveys



(a) (b)

Fig. 5. Correlograms of feature num_restless_bout (number of restless periods in sleep) in time window 4 for two students (left: a student in L_Pre1_Post1, right: a student in L_Pre1_Post2).

- Pre1_ Post2: not having a mental health condition in the pre-semester survey, but having it in the post-semester survey
- Pre2_ Post2: having a mental health condition in both surveys
- Pre2_ Post1: having a mental health condition in the pre-semester survey, but not in the post-semester survey

The following sections describe our observations and findings. To distinguish the mental health groups in the two conditions, we add an *L* and *D* to the mental health group for loneliness (e.g., L_Pre1_Post2) and depression (e.g., D_Pre1_Post2), respectively.

4.1.3 Detection of Rhythmicity and Regularity in Student Data. To investigate whether rhythmicity exists in data collected from students' smartphones and Fitbits (Question 1) and whether students' rhythms remain stable throughout the semester (Question 2), we used Autocorrelation and Fourier Periodogram to model students' rhythms in each time window for each sensor feature.

We first applied the Autocorrelation on a sleep feature which indicates that students with high loneliness have less stable sleep rhythms. Figure 5 shows the correlogram of the number of restless sleep bouts in two students from different groups, one with low loneliness throughout the semester and the other with high loneliness at the end of the semester. The figure visually depicts differences in the rhythms of these two students where the correlogram belonging to the student with high loneliness projects a less stable rhythm towards the end of the time series. To further quantify such differences in cyclic rhythms of students, we applied Periodogram to (1) detect dominant periods in students' data, and (2) measure variability in those periods among students with different health statuses.

To identify the dominant periods, the Fourier periodogram is used to detect all significant periods for each sensor feature. The results of the periodogram show that the most dominant cyclic periods in each time window are 24- and 12-hours for all sensor features. For example, for sleep duration feature in the depression category, this trend is consistent in all students regardless of the mental health condition where on average 97.6% and 69.6% of students have 24- and 12-hours as dominant periods in their data across time windows (Tables 1 and 2). The percentages, however, have a declining trend starting from TW4 (around midterms) towards the end of the semester. This trend can be expected because of the increase in students' workload that causes irregularity in sleep duration. The lowest percentages across all time windows (46.3% on average) are observed in the 12-hour period of students in group D_Pre2_Post2, i.e., students who were depressed throughout the semester. In particular, there is no 12-hour period observed for this group in TW1 (the first two weeks) and TW8 (the last two weeks). The 12-hour or half-day period relates to diurnal/nocturnal activities, and this trend may be indicative of higher irregularity in sleep behavior

Table 1. Top Two Dominant Periods of Sleep Duration Feature for Depression Groups

	TW1			TW2			TW3			TW4			TW5			TW6			TW7			TW8		
Group	N	P1 (%)	P2 (%)	N	P1 (%)	P2 (%)	N	P1 (%)	P2 (%)	N	P1 (%)	P2 (%)	N	P1 (%)	P2 (%)	N	P1 (%)	P2 (%)	N	P1 (%)	P2 (%)	N	P1 (%)	P2 (%)
All	125	24 (98)	12 (70)	120	24 (98)	12 (75)	118	24 (95)	12 (71)	115	24 (88)	12 (51)	104	24 (88)	12 (52)	103	24 (88)	12 (65)	101	24 (89)	12 (53)	97	24 (94)	12 (69)
D_Pre1_Post1	72	24 (97)	12 (69)	68	24 (99)	12 (72)	66	24 (98)	12 (74)	67	24 (85)	12 (46)	60	24 (87)	12 (52)	58	24 (90)	12 (66)	57	24 (88)	12 (51)	58	24 (98)	12 (72)
D_Pre1_Post2	35	24 (100)	12 (89)	34	24 (97)	12 (88)	35	24 (91)	12 (80)	33	24 (91)	12 (64)	33	24 (97)	12 (58)	33	24 (94)	12 (64)	33	24 (91)	12 (61)	28	24 (93)	12 (79)
D_Pre2_Post1	2	24 (100)	12 (100)	2	24 (100)	12 (100)	2	24 (100)	12 (50)	2	24 (100)	12 (100)	1	24 (100)	31.2 (100)	1	24 (100)	12 (100)	1	24 (100)	12 (100)	2	24 (100)	12 (50)
D_Pre2_Post2	16	24 (94)	156 (38)	16	24 (94)	12 (56)	15	24 (87)	12 (40)	13	24 (92)	12 (38)	10	24 (70)	12 (40)	11	24 (64)	12 (64)	10	24 (90)	12 (40)	9	24 (67)	54 (33)

N is the number of students in the group. P1 is the most dominant period (i.e., the percentage of students that have the period is highest among all periods). The percentage in parenthesis is the percentage of students with that period. P2 is the second dominant period.

Table 2. Top Three Dominant Periods of Sleep Duration (minutes asleep) Feature for Pre1_Post2 Groups

	Pre1_Post2							
	Loneliness				Depression			
Time Window	N	P1 (%)	P2 (%)	P3 (%)	N	P1 (%)	P2 (%)	P3 (%)
TW1	17	24 (100)	12 (71)	312 (35)	35	24 (100)	12 (89)	312 (34)
TW2	15	24 (93)	12 (87)	312 (40)	34	24 (97)	12 (88)	312 (38)
TW3	16	24 (100)	12 (88)	156 (31)	35	24 (91)	12 (80)	156 (31)
TW4	15	24 (73)	12 (53)	312 (33)	33	24 (91)	12 (64)	78 (40)
TW5	14	24 (100)	12 (64)	156 (29)	33	24 (97)	12 (58)	312 (36)
TW6	12	24 (92)	12 (67)	78 (33)	33	24 (94)	12 (64)	78 (45)
TW7	13	24 (85)	12 (54)	156 (31)	33	24 (91)	12 (61)	156 (40)
TW8	11	24 (91)	12 (55)	72 (45)	28	24 (93)	12 (78)	72 (32)

N is the number of students in the group. P1 is the most dominant period (i.e., the percentage of students that have this period is highest among all periods). The percentage in parenthesis is the percentage of students that have the period. P2 and P3 are the second and third dominant periods.

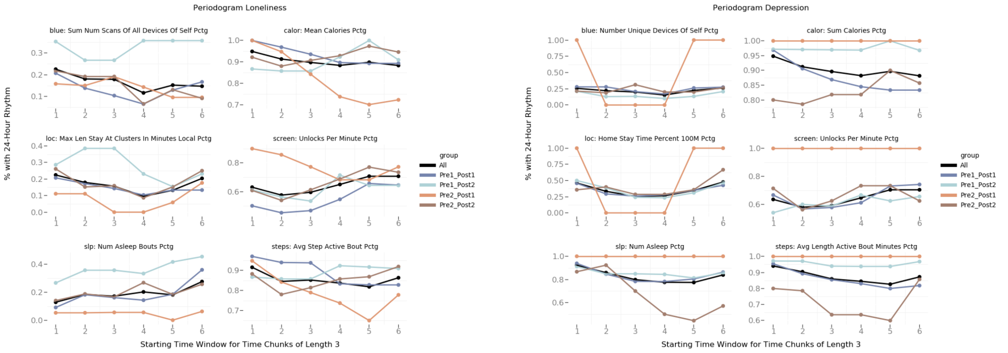
among students with depression throughout the semester especially at the beginning and towards the end of the semester. Our observations are consistent with other studies. [51] observed that older adults with depression have a lower sleep regularity index in a study of 138 participants. [62] observed that irregular sleepers showed more negative moods, including depression, in a study of male college students.

We picked the sleep duration to further analyze changes in periodicity in students who started the semester with normal health status but developed depression or loneliness towards the end (D_Pre1_Post2 or L_Pre1_Post2). Table 2 shows that the dominant periods of 24- and 12-hours are preserved for the sleep duration feature in all time windows for both loneliness and depression groups. While the same declining trend towards the end of the semester exists for both loneliness and depression groups, a sharper slope is observed for the 12-hour period. The lowest percentage of students in this group with 24- and 12-hour periods are in time windows 4 and 5 with 73% in loneliness category (24-hour), 91% in depression category (24-hour), 53% in loneliness category (12-hour), and 57% in depression category (12-hour). Given that time windows 4 and 5 intersect with midterm and spring break, these observations point to changes in sleep patterns among students whose mental health worsens over the semester.

The third dominant periods for sleep duration across all time windows include 312-hour (13 days), 156-hour (6.5 days), and 78-hour (3.25 days). This is an interesting observation as these numbers are multiplies of the 78-hour period. In other words, it seems the sleep duration of roughly one third of the population in these groups follows a weekly pattern that may be imposed by class schedules.

Table 3. Percentage of Participants with 24-hour Period Across all Sensor Features

% of Participants with 24-hour period										
Audio	Battery	Bluetooth	Calorie	Location	Location Map	Call&Messages	Screen	Sleep	Steps	Wifi
62	13	42	92	41	17	18	36	69	95	83



(a) (b)

Fig. 6. The plots show the percentage of participants with 24-hour as the dominant rhythm (y-axis) in each mental health group (left: loneliness, right: depression) for each time chunk of length 3 (x-axis). The data point at $x = i$ corresponds to the time chunk of length 3 starting at tw_i (i.e., tc_{3i}). It represents the percentage of participants with 24-hour as the dominant rhythm in all the 3 time windows tw_i, tw_{i+1}, tw_{i+2}.

Overall and across all sensor features, we observe the 24-hour as the dominant period for over 52% of the student population with the highest percentages belonging to steps (95%), calories (92%), Wi-Fi (83%), and sleep (68%). Table 3 presents the overall percentages for each sensor. Calories and steps relate to physical activity. The high percentage of students with 24-hour cycles in these two sensor categories is indicative of regular daily exercise and movement. While there is a low percentage of students with regularity in their cyclic location patterns and visited places (Location Map features), it seems a large number of students have regular daily patterns of using Wi-Fi. This pattern could be expected given that the first-year students live in dorms and are mostly on campus. Interestingly, a low percentage of students seem to have regular cyclic patterns of phone usage (Screen, 36%; Call & Messages, 18%; Battery 13%). While phone use especially battery charging patterns are expected to be cyclic (e.g., charging the phone at night), these observations present the possibility of different phone use behavior among students.

To measure the variability of the dominant periods among students with different health statuses, we look at the percentage of participants in each mental health group that had 24-hours as one of their dominant rhythms for each *time chunk*. This would help observe the extent to which students preserved their normal circadian rhythm over the semester. Recall that time chunks consist of k consecutive time windows, there were 36 different time chunks in total for eight time windows of length 2 in the dataset. In each time chunk, a participant had 24-hour as a dominant rhythm if and only if this participant had 24-hour as a dominant rhythm in all time windows in that time chunk. Figure 6 shows the percentage of participants with 24-hour as the dominant rhythm (y-axis) in each mental health group for each time chunk of length 3 (x-axis). We chose one representative feature from each sensor stream, i.e., Bluetooth (abbreviated as blue in the figure), location (loc), sleep (slp), calories (calor), screen, and steps for further analysis. As shown in Figure 6, the trend in the percentage of 24-hour rhythms varies a lot in mental health groups

Table 4. The Table Lists the Aggregated Variance of the Percentage of 24-hour Rhythms Across Time Chunks for Loneliness and Depression Separately

Mental Group	all	pre1_pre1	pre1_pre2	pre2_pre1	pre2_pre2
Loneliness	0.04	0.05	0.06	0.07	0.05
Depression	0.05	0.05	0.05	-	0.09

We first calculated the variance per mental health group in each sensor feature shown in Figure 6, and then averaged these variance values across sensor features of loneliness or depression. The aggregated variance can represent the stability of rhythms of each mental health group.

and across time chunks in each sub-figure. To understand the significance of these variations, we 1) applied K-W ANOVA (Kruskal-Wallis one-way analysis of variance) [12] to test the variance of trends across mental health groups, and 2) calculated the variance in the percentage of 24-hour rhythms for each mental health group across time chunks. For loneliness, the trends for all features show significant differences among mental health groups (the average/median of p-value across sensor features is 0.02/0.03). For depression, mental health groups have more similar trends. In contrast to Bluetooth, calorie, and step features that have significant differences in their trends (p-values of 0.05, 0.001, and 0.001), location, sleep, and screen features do not show any significant differences (p-values 0.94, 0.26, and 0.67). This is visually demonstrated in Figure 6, e.g., the trend for location is similar for all four depression groups. We also calculated the average variance for each mental health group across sensor features. As shown in Table 4 for loneliness, most changes in the 24-hour rhythms were observed in the group with high loneliness at the beginning and low loneliness at the end of the semester (pre2_pre1) group whereas for depression, the group with depression throughout the semster (pre2_pre2) had the largest fluctuations.

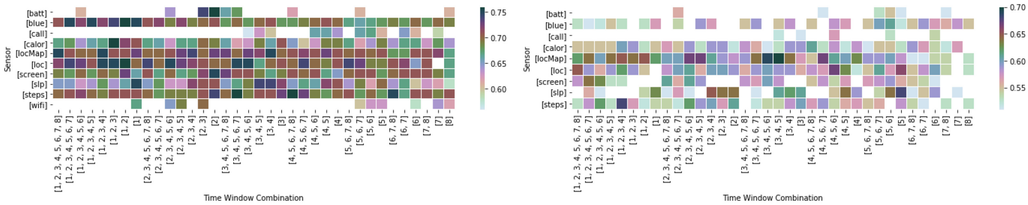
For loneliness, the group with low loneliness at the beginning and high loneliness at the end of the semester (L_Pre1_Post2) shows an overall higher percentage of 24-hour rhythms for features of sleep, location, and Bluetooth across time windows. The opposite group with high loneliness at the beginning and low loneliness at the end of the semester (L_Pre2_Post1) shows a lower percentage of 24-hour rhythms for features of calories and steps but higher percentages for screen features. The Bluetooth feature in the top left of Figure 6(a) which represents the cyclic patterns of the scanned devices belonging to the person is a proxy of social isolation, i.e., the person not being around other people (and their devices) and being mostly by themselves. Starting from TW3 (week 3, 4, and 5), the percentage of students with regular daily cycle for this features in L_Pre1_Post2 and L_Pre2_Post1 groups sharply increase and decrease, respectively. In other words, while more students with low loneliness at the beginning and high loneliness at the end of the semester start having a regular social isolation pattern on a daily basis towards the end of the semester, fewer students in the opposite group with high loneliness at the beginning and low loneliness at the end of the semester experience this trend. A very similar pattern is observed for another socially relevant feature namely the length of stay in significant locations. The trend is relatively stable and slightly decreasing in students with no loneliness which reflects the stability of behavior in this group. For sleep, steps, and calorie burn, we observe an almost counterintuitive opposite cyclic behavior among L_Pre1_Post2 and L_Pre2_Post1 groups. It seems more students with loneliness toward the end of the semester engage in regular physical activities as projected by calories and steps features and have more regular sleep duration cycles. A relatively similar behavior is observed for the burned calories feature in depression groups (Figure 6 top right). While regularity in physical activities slightly increases in students with depression (D_Pre2_Post2), it appears to decrease in students with no depression (D_Pre1_Post1) across time windows. While existing studies, e.g., [7, 20, 61] point to negative associations of physical activities and mental health, we believe the increase in regular physical activities towards the end of the semester may be a coping attempt by students with mental health problems.

But trends generally look different for depression groups in Figure 6(b). All groups except D_Pre2_Post1 had similar percentages of regular 24- and 12-hour periods for Bluetooth, location, and screen across time windows. While the group with no depression at the beginning and with depression at the end of the semester (D_Pre1_Post2) shows the highest percentage of normal 24-hour rhythms for features of calories and steps across all time windows, the group that was depressed throughout the semester (D_Pre2_Post2) shows the lowest percentages for steps, sleep, and calories. In particular, the regularity of sleep in these students seems to decline drastically across time windows. Although expected, this sharp trend is a valuable observation for further exploration of relationships between change in sleep cycles and depression status. The previous study in [51] also observed that sleep irregularity is indicative of depression, but no existing study has analyzed the relationship between change in sleep cycles and change in depression status. Our observations provide new findings and insights that call for further and more rigorous investigations.

4.1.4 Prediction of Mental Health Status with Rhythmic Features. The third and fourth questions in our analysis relate to the feasibility of using biobehavioral rhythm parameters to predict students' mental health status. In our framework, we utilize dominant periods that were detected using Fourier Periodogram described in Section 4.1.3 to build Cosinor models of biobehavioral data. This process generates rhythmic features fed into the machine learning process to classify post-semester loneliness and depression categories (low loneliness vs. high loneliness and no depression vs. with depression) of the students. We build two types of datasets, one with single sensors only and one with multiple sensors. In the following paragraph, we will evaluate the performance of single sensor modeling and multiple sensor modeling to find out what types of sensor features and rhythmic features contribute most to the prediction.

For *Single Sensor* datasets, we use the rhythmic features of each sensor feature separately, i.e., for each sensor feature and each time window (with time windows of two weeks), we take the rhythmic features of this sensor feature and time window to form the input dataset. We remove datasets with more than 30% missing instances (80 training instances) as we consider it too small to generate a reliable and generalizable model. For *Multiple Sensors* datasets, we select the sensor features that provide accuracy above baseline in models built with single sensors. For both approaches, we use the majority class ratio i.e., the category that has the highest percentage of labels for that category as the comparison baseline. We then repeat the same process we followed for single sensor datasets, but this time for the combination of sensor features, i.e., for each combination of sensor features and each time window, we take the rhythmic features of the selected sensor features of those sensors and time window to form the input datasets. Other than the difference in the input dataset, the machine learning pipeline is the same for the two types of datasets.

Given the imbalanced datasets for both health conditions i.e., the different number of samples in the two classes (e.g., 59% of samples in category 1 vs. 41% in category 2 of depression), using the accuracy will not be adequate for performance evaluation and needs to be accompanied by other measures such as F1. For every combination of time window and sensor, the F1 score is used to select the model with the best performance. We build models with single sensor and multiple sensors datasets for both mental health conditions. The results of all combinations are shown in Figures 7 and 8. The heatmaps use the depth of color to represent the F1 score. Given a large number of features, we only report results with accuracy above the baseline (majority class percentage). Through the single sensor modeling, we can judge which type of sensor is most effective in predicting mental health. Overall, we find that the models with multiple sensors improve the prediction performance. A summarization of the results are listed in Table 5.



(a) Loneliness

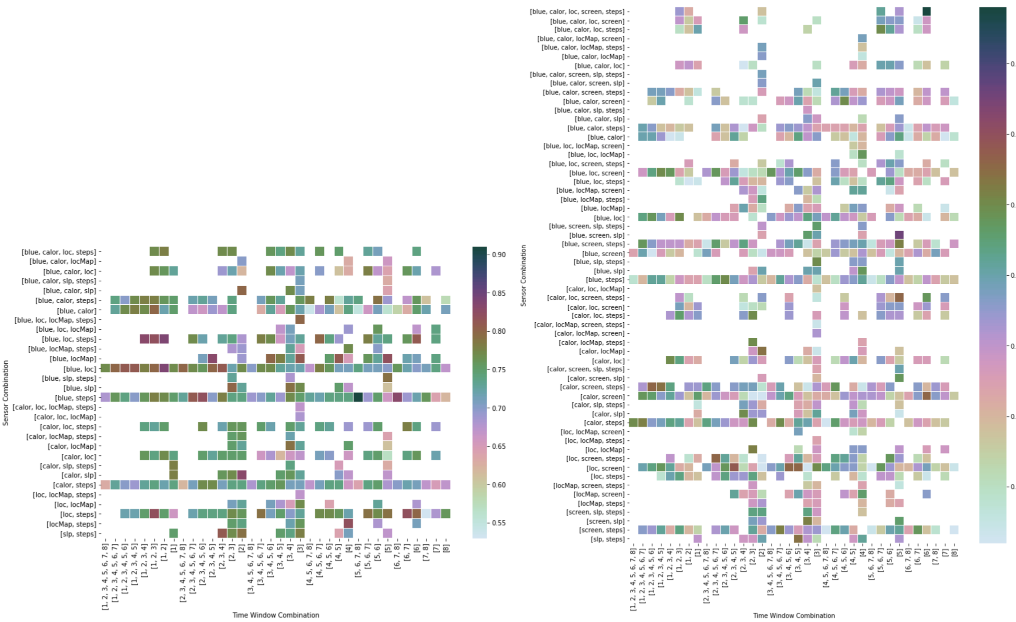
(b) Depression

Fig. 7. The heatmap displays the largest F1 score in the loneliness and depression prediction model trained by a combination of different single sensor features and time windows.

Single Sensor Modeling. The F1 scores of machine learning models with single sensor features are shown in Figure 7. Overall, the models for loneliness prediction obtain higher accuracy (F1) scores than depression models (Table 5) which may be due to more sparsity in depression datasets. Rhythm parameters obtained from Cosinor models built for features related to Bluetooth, calories, location, sleep, and steps perform better in predicting both loneliness and depression levels. Although the best model to classify post-semester loneliness is built using Gradient Boosting on rhythm parameters of calorie data from tw_1 to tw_3 with an F1 score of 0.76, more models built on rhythms of location and locationMap provide high performance. The best model for post-semester depression with an F1 score of 0.7 is also built using Gradient Boosting but on the locationMap data from tw_3 to tw_5. Compared to other sensors, models using rhythmic parameters from locationMap features show better performance for predicting post-semester depression (six out of ten models with the highest F1 score use locationMap features). Although the F1 scores of models with a single time window are generally lower than models with multiple time windows, there are some exceptions in the heatmaps of both loneliness and depression. For example, the loneliness model using sleep features in tw_1 achieves an F1 score of 0.75, and the F1 score of the depression model using sleep features in tw_5 equals 0.68. Interestingly and somewhat counter-intuitively, across all sensors, the majority of models (avg. 57.5% for single sensors and 53.5% for multiple sensors) using early semester time windows (tw_1 to tw_4) appear to have higher F1 scores for post-semester loneliness and depression prediction than late semester time windows. We believe this observation provides initial evidence for the possibility of early detection of mental health status via monitoring of changes in biobehavioral rhythms.

Multiple Sensor Modeling. We do the same analysis for the combination of sensor features. From Figure 8, we observe that the combination of multiple sensor features contributes to the improvement of the F1 score. For example, the combinations related to steps, sleep, location, calorie, and Bluetooth end with better results. For predicting loneliness, the best model is built with Logistic Regression, which uses the Bluetooth and steps data from tw_5 to tw_8 and obtains an F1 score of 0.91. For predicting depression, the best model is obtained from Logistic Regression using the rhythm parameters from Bluetooth, calorie, location, screen, and steps features. The model only uses tw_6 to predict depression with an F1 score of 0.89. The best model predicting depression has a lower F1 score than the best model predicting loneliness, which is the same as the single sensor model and may be due to sparsity in sensor data.

Table 5 summarizes the mean and max of F1 scores for models built with each combination of the feature selection and machine learning methods. In single sensor modeling, the combinations of Logistic Regression with Lasso and Randomized Logistic Regression perform best for predicting loneliness with the mean and max F1 score of 0.7 and 0.76, respectively. The combination of Gradient Boosting and Information Gain provides the highest F1 score for the prediction of depression.



(a) Loneliness

(b) Depression

Fig. 8. The heatmap displays the largest F1 score in the loneliness prediction model trained by a combination of different multiple sensor features and time windows.

Table 5. Summary of the Mean and Maximal Values of F1 Scores for Each Combination of Feature Selection and Machine Learning Methods Shown in the Heatmaps 7, 8

	Single Sensor						Multiple sensors					
	Loneliness mean(max)			Depression mean(max)			Loneliness mean(max)			Depression mean(max)		
	GB	LR	RF	GB	LR	RF	GB	LR	RF	GB	LR	RF
IG	**0.69 (0.76)**	**0.69 (0.76)**	0.66 (0.72)	**0.58 (0.70)**	**0.60 (0.61)**	0.56 (0.63)	0.73 (0.83)	0.72 (0.78)	0.69 (0.81)	0.63 (0.83)	0.60 (0.66)	0.63 (0.76)
Lasso	0.68 (0.72)	**0.70 (0.76)**	**0.74 (0.74)**	0.57 (0.68)	0.57 (0.64)	0.55 (0.59)	0.72 (0.78)	**0.75 (0.91)**		0.59 (0.66)	**0.67 (0.89)**	0.54 (0.54)
RLR		**0.70 (0.76)**	0.68 (0.73)	0.58 (0.65)	0.56 (0.65)	0.57 (0.60)	0.75 (0.81)	0.73 (0.82)	**0.76 (0.84)**	0.65 (0.78)	0.65 (0.79)	0.65 (0.79)

The bold values are either the biggest mean value of F1 scores, or the biggest maximal values of F1 scores.

For the multiple sensor modeling, we observe that the maximum F1 scores of predicting loneliness and depression are 0.91 and 0.89, which are obtained from the combination of Logistic Regression and Lasso. Overall, for the majority of approaches, the combination of Gradient Boosting and Information Gain provides the best performance. This combination should be further evaluated with other similar datasets to replicate and confirm their superior performance over other algorithm combinations.

Dominant rhythm parameters that predict mental health. We count the frequency of rhythmic features selected by machine learning models to measure the contribution of each rhythm parameter in predicting mental health. Orthophase and Magnitude appeared on top of the list as the most frequently selected parameters. Although we used three feature selection methods in our evaluation, we observed that the Information Gain method provided a more reliable and complete list of features during the training. Table 6 shows the rhythm features that are selected most frequently by Information Gain during depression prediction for each sensor feature in each time window. The **vertical dominant feature (VDominant)** is the most commonly selected feature for most of the sensors at a given time window, and the **horizontal dominant feature**

Table 6. The Most Frequently Selected Rhythmic Features by Information Gain During Depression Prediction

	TW1	TW2	TW3	TW4	TW5	TW6	TW7	TW8	HDominant
Audio	Amp SE	Mesor SE	Amp SE	IPR	Magnitude	Amp SE	Bathyphase	P	Amp SE
Battery	IPR	PR	Mesor SE	Mesor SE	Orthophase	Magnitude	Orthophase	Bathyphase	Mesor SE
Bluetooth	Magnitude	Bathyphase	Amp	P	IPR	Orthophase	Mesor SE	Orthophase	Orthophase
Call	IPR	PHI	IPR	IPR	Amp SE	Bathyphase	Orthophase	Magnitude	IPR
Calorie	Mesor	Magnitude	Magnitude	Bathyphase	Orthophase	Orthophase	IPR	Magnitude	Magnitude
Location	PHI SE	Magnitude	Mesor	PR	IPR	Mesor	Amp SE	IPR	Mesor
Location Map	Orthophase	Magnitude	Mesor	Orthophase	PHI	Bathyphase	Orthophase	Bathyphase	Orthophase
Messages	Orthophase	Magnitude	LCM	PR	Mesor SE	Bathyphase	PHI SE	Magnitude	Magnitude
Screen	Amp	P	Orthophase	Orthophase	PR	Orthophase	IP	Amp SE	Orthophase
Sleep	Bathyphase	PHI SE	Mesor	Orthophase	PHI SE	IP	Amp SE	Bathyphase	Bathyphase
Steps	P	Orthophase	Magnitude	Bathyphase	PR	IPR	IPR	Magnitude	Magnitude
Wifi	Amp	Mesor SE	Mesor	Orthophase	Magnitude	IPR	IP	Amp SE	Magnitude
VDominant	Amp	Magnitude	Mesor	Orthophase	Orthophase	Bathyphase	Orthophase	Magnitude	**Orthophase**

Table 7. F1 of Machine Learning Models with Rhythm Modeling (rhythm) and Without Rhythm Modeling (raw features)

Time Window	Feature	Rhythm-F1	Raw-F1	Time Window	Feature	Rhythm-F1	Raw-F1
1	shortest period spent at Halls	0.66	0.54	1	shortest period spent at Halls	0.69	0.55
2	longest awake period length	0.64	0.49	2	longest awake period length	0.67	0.47
3	number of awakes	0.63	0.47	3	total asleep time	0.67	0.49
4	maximum calories increase between 5-min periods	0.66	0.60	4	number of awakes	0.62	0.56
5	shortest asleep period length	0.70	0.69	5	percentage of time spent moving	0.72	0.52
6	total distance traveled	0.65	0.50	6	longest period spent at athletic areas	0.68	0.43
7	maximum calories decrease between 5-min periods	0.67	0.59	7	total change of calories	0.68	0.53
8	minutes spent at Halls	0.65	0.62	8	variance of moving speed	0.67	0.48

Left: Loneliness; Right: Depression.

(HDominant) is the most commonly selected feature in most time windows for a given sensor. The overall dominant feature (the feature at the bottom right corner in bold font) is the most commonly selected feature for all sensors and time windows. If two features are the most commonly selected features for the same number of sensors/time windows, we break the tie by taking the feature with a higher frequency. Overall, Orthophase is selected most frequently for all sensors and time windows. Magnitude comes in second. Given that Phase and Magnitude reflect duration and intensity of biobehavioral features, frequent selection of these parameters suggests an important relationship with mental health status.

In addition to the main rhythmic features, i.e., Mesor, Amplitude/Magnitude, and Ortho/Bathyphase, we observe frequent selection of features related to the fit of Cosinor models including the significance level of the fit (P), Standard Errors (SE) and Percent Rhythm (PR and IPR), i.e. the proportion of the overall variance accounted for by the fitted model. Higher levels of these parameters reflect higher variation in data. Therefore, frequent selection of these parameters indicates the power of regularity/irregularity of biobehavioral rhythms in predicting mental health status.

4.1.5 Comparison with Models Built Without Rhythm Parameters. To better understand the capability of our framework in utilizing rhythmic features to predict an outcome, we compare the prediction performance of the models with rhythm modeling against the models without rhythm modeling. Specifically, we select the best performing sensor feature in each time window, run exactly the same machine learning pipeline on the raw feature data without rhythm modeling, and compute the F1 score. Table 7 shows that the pipeline with rhythm modeling outperforms the one without by a large margin on most of the features. This observation is consistent with both loneliness and depression predictions.

4.2 Case 2: Biobehavioral Rhythm Modeling for Readiness Prediction Using Data from OURA Ring

We chose a second dataset to evaluate the framework's flexibility in modeling various types of data and applying different analysis approaches. The sensors, participants, and ground truth of the dataset used in case 2 are different from case 1. For this case, we used data from 11 undergraduate and graduate students who continuously wore the OURA ring, a commercially available smart and convenient health tracker, for several months.

As shown in the last plot of Figure 10, the length of data collection varies per participant and ranges from 31 to 323 days. The long-term data makes it possible to detect and observe rhythms with larger cyclic periods than a day, e.g., weeks or months. As such, we design our analysis to answer the following:

(1) Are there common cycles in participants' data per sensor and across sensors, and can we identify similarities and differences in cyclic periods among participants despite differences in the length of their data?
(2) How accurately can individual rhythm models per sensor feature and per participant predict average readiness?

4.2.1 Physiological Data Processing. OURA collects sleep, heart rate, skin temperature, calories, steps, and activity. Sleep, heart rate, and skin temperature samples are collected every five minutes during night hours; and activity, calories, and steps are sampled every five minutes during the day. The data is summarized and stored on the OURA cloud platform. As our goal is to detect cycles with multiple-day lengths, we aggregate the features into daily intervals (as opposed to the previous case that used hours). In total, we use 31 features such as total duration of sleep, lowest/average heart rate, average metabolism level, total amount of calories burned, and total number of steps during the day. To be able to detect the longest periods in participants' data, we refrain from segmenting data into common time windows and use the entire time series data for the analysis. The convenience of wearing the ring and its long battery life lead to good quality data with low missing rates (Max 15.6% in our data). We use the moving average method to handle the missing values.

4.2.2 Readiness Score as Ground Truth. Besides the physiological features, OURA provides a readiness score displayed in the OURA ring app, i.e., an evaluation of the body's overall recovery rate after waking up in the morning. The readiness score ranges from 0 to 100 with scores over 85 indicating high readiness for challenging tasks and scores below 70 indicating poor body state and need for recovery. In our dataset, participants' readiness scores range from 24 to 99 with an average score of 74, and a standard deviation of 11.4. Figures 10 and 9 shows the distribution and variation of daily readiness score for each participant. We calculate the average daily readiness score for each participant and use it as ground truth to explore how well we can use the rhythms to predict the readiness score.

4.2.3 Detection of Cycles in OURA-Ring Data. Our first analysis questions relate to detecting common cycles in the participants' data and physiological sensors. Our results show weekly and biweekly periods were observed most frequently. Similar to case 1, we applied Fourier Periodogram on the time series data of each sensor feature per participant to detect significant periods. In Tables 8 and 9, we list the most frequently detected periods of sensor features and summarize them by sensor type and participants. The number 7 and its multiple 14 as well as its close preceding and following numbers of 6 and 8 appear most frequently in both tables suggesting near-weekly biobehavioral patterns. In particular, periods of Activity, Sleep, and Heart

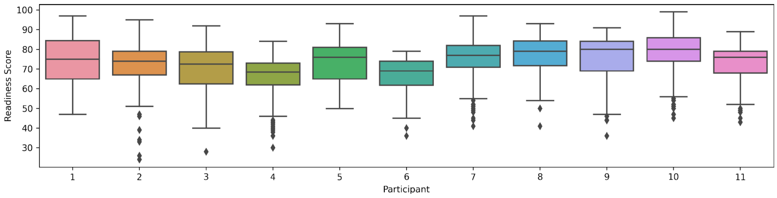


Fig. 9. The 1 to 11 boxplots display the minimum, median, maximum, and quartile of the daily readiness scores for each participant. Most daily readiness scores are clustered in the range from 70 to 85.

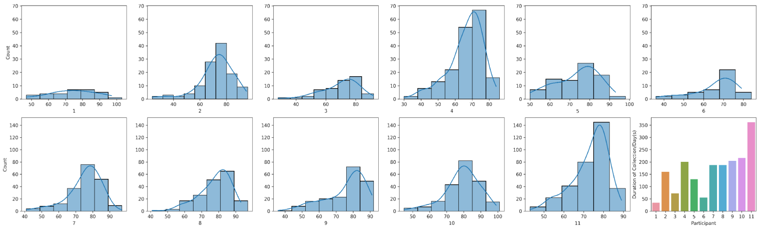


Fig. 10. The histograms from 1 to 11 display the distribution of the daily readiness scores for each participant, and the last bar plot shows the duration of each participant's data collection.

Table 8. Dominant Frequent Periods for Each Sensor

Sensor	Detected Period (% of Participants)
Activity	7 (55), 2 (45), 6 (45), 8 (36), 4 (36)
Calorie	2 (18), 11 (18), 10 (18), 4 (9), 81 (9), 20 (9)
Heart Rate	7 (36), 27 (27), 8 (27), 14 (18), 18 (18)
Sleep	8 (55), 3 (55), 7 (45), 6 (45), 11 (36)
Steps	11 (27), 10 (27), 2 (18), 54 (18), 7 (18)
Skin Temperature	12 (36), 14 (36), 15 (27), 27 (27), 34 (18)

The percentage in parenthesis is the percentage of participants with the significant period.

rate project near-weekly cycles across all participants. For example, Activity cycles of 6, 7, and 8 days are observed in 45%, 55%, and 36% of participants, respectively. These cycles are also observed in sensor data of seven participants (63%). Calorie and Steps share periods of 2, 10, and 11 days with similar percentages. Although the percentages of participants with these cycles are low likely due to different movement patterns among participants, the common periods of these two sensors may be indicative of exercise cycles in those participants.

4.2.4 Prediction of Readiness with Rhythmic Features. For each participant, we use the three most significant periods identified by the Periodogram as input to the Cosinor method to build rhythm models per sensor feature. The rhythmic features are then entered into the machine learning process to predict average readiness per participant. Since the readiness score is a continuous variable, we build regression models to make predictions. Our choice of machine learning algorithms includes Random Forest and Gradient Boosting with Information Gain and Lasso as feature selection methods. Similar to case 1 in mental health, we build models with single and multiple sensor combinations in a leave-one-participant-out cross validation, but, instead of accuracy, we use the **Root Mean Square Error (RMSE)** as the performance measure.

Table 9. Most Frequent Periods of all Sensor Features for Each Participant

Participant	Detected Period (% of Sensor Features)
1	7 (29), 34 (26), 2 (23), 3 (16), 39 (10)
2	80 (42), 81 (39), 40 (35), 11 (29), 32 (26)
3	77 (32), 10 (29), 24 (23), 7 (23), 26 (16)
4	7 (52), 202 (39), 101 (35), 67 (19), 201 (16)
5	66 (39), 65 (35),130 (26), 8 (26), 26 (23)
6	6 (35), 56 (29), 14 (23), 28 (13), 19 (10)
7	31 (26), 11 (23), 190 (23), 95 (23), 38 (19)
8	94 (42), 188 (29), 63 (29), 7 (23), 189 (23)
9	68 (45), 102 (35), 29 (29), 204 (26), 41 (16)
10	54 (45), 108 (39), 43 (35), 27 (23), 217 (32)
11	126 (35), 42 (26), 28 (23), 5 (16), 7 (16)

The percentage in parenthesis is the percentage of sensor features with that period.

Table 10. Lowest RMSE of Single Sensor Features and Frequent Rhythmic Features Selected by IG and Lasso

Sensor	Activity		Calorie		HR		Sleep		Step		Skin Temperature	
Feature Selection	IG	Lasso	IG	Lasso	IG	Lasso	IG	Lasso	IG	Lasso	IG	Lasso
RMSE (GB)	5.04	8.42	4.79	5.18	4.54	5.50	4.08	5.54	4.71	6.77	5.34	6.77
RMSE (RF)	5.25	8.52	4.38	4.51	4.65	6.20	4.20	5.68	4.81	7.30	5.48	7.30
Frequent Rhythmic Features	PR	PHI	Mesor SE, Amp SE	PHI	PR	PHI, PHI SE, P	PR	P	Mesor SE, Amp SE	Mesor, P	Mesor SE, Amp SE, P	PHI

Table 10 lists the best RMSE achieved by single sensor models along with the most frequently selected features. Among single sensor models, the model built with the rhythmic feature of sleep data with an RMSE of 4.08 is a stronger predictor of readiness than others. In comparison, the combination of sleep, calories, and steps obtain an RMSE of 3.54, the lowest RMSE among all multiple sensor models, as shown in Table 11. This combination considers both the activity of the human body during the day (calories) and the sleep quality at night (sleep). These observations are expected and confirm the impact of both sleep and physical activity on the body's daily functioning. Interestingly but not surprisingly, the frequently selected features across all sensors are standard errors of the rhythm parameters (i.e., PHI SE, MESOR SE, and Amp SE) as well as percent rhythm (PR), all of which are indicative of variation in the actual data. MESOR SE is the most dominant feature among both single and multiple sensor models. These results suggest that the level of variability and potentially irregularity in biobehavior may be most predictive of fluctuations in readiness.

Tables 10 and 11 also summarize the RMSE for models using each combination of feature selection and machine learning methods. The Gradient Boosting model with Lasso regression achieves the best performance for both single sensor and multiple sensor modeling, with an RMSE of 3.54. Using the same prediction model, the Information Gain performs better in single sensor modeling, and the results are reversed in multiple sensor modeling.

5 DISCUSSION

In the Introduction section, we identified several challenges in processing and modeling biobehavioral time series data from mobile and wearable devices that motivated the development of our novel computational framework. These challenges include 1) automated handling and processing of massive multimodal sensor data, 2) granular and fine-grained exploration of all signals to

Table 11. RMSE of Multiple Sensor Models and Frequent Rhythmic Features Selected by Those Models

Feature Selection	IG	Lasso
Sensor	sleep, calorie, step	sleep, calorie, step
RMSE (GB)	3.73	3.54
RMSE (RF)	3.80	3.68
Frequent Rhythmic Features	MESOR SE	MESOR

extract knowledge about biobehavioral cycles, and 3) computational steps for modeling, discovering, and quantification of common patterns.

We presented two case studies using different datasets, sensors, populations, and prediction tasks to demonstrate the capabilities of our proposed computational framework in addressing the aforementioned challenges. Both cases demonstrated the ability of the framework to automatically process longitudinal multimodal sensor mobile data; extract fine-grained and granular features; detect periodicity in the data and use it to study rhythm stability and variation over time; build micro-rhythm models for each biobehavioral feature; and use those models to incorporate different analytic approaches to predict various health outcomes. We were able to build massive prediction models for both single sensors and different combinations of sensors and to compare the results. We observed that the combination of multiple sensor features contributed to the improvement of prediction results. We also showed that the models built with rhythmic features outperform models built with the raw sensor features further demonstrating the feasibility of biobehavioral rhythms in prediction tasks.

Although our primary goal was to showcase the capabilities and flexibility of the framework, our analyses provided interesting and novel observations, some of which can be used as initial evidence for further investigation. For example, although we used different datasets and population groups in cases 1 and 2, we observed near-weekly sleep cycles in both populations. We also observed a drastic decline in sleep duration cycles of depressed students throughout the semester. Even though existing research has repeatedly shown relationships between sleep and mental health, we believe our observation is unique in identifying relationships between change in cyclic patterns of sleep and mental health status. Our micro machine learning models of sensor features provided evidence that changes in biobehavioral rhythms in the early weeks of the semester were predictive of post-semester depression and loneliness. This finding suggests monitoring biobehavioral rhythms may serve as a useful tool for early prediction of change in mental health status. We also observed that rhythmic parameters of Phase and Magnitude that reflect duration and intensity of biobehavioral features as well as parameters related to variability in the cyclic time series models (e.g., SEs and PR) were frequently selected in the machine learning process indicating the power of the intensity, duration, and regularity/irregularity of biobehavioral rhythms in the prediction of health outcomes. Since there is no comparable study in biobehavioral rhythms for the prediction of health and wellness, we only compared our observations with the closest studies of loneliness and depression. We hope our initial findings opens up for more studies using our framework to replicate the results.

The central theme of this paper was introducing the computational framework and its main functionality. However, the framework can be adapted and extended to include more functionalities and features. The advancements include 1) adding more data sources such as weather, environment, work schedules, and social engagements to draw a more holistic picture of biobehavioral rhythms in individuals and groups of people, 2) adding a conclusive set of periodic functions and methods with diverse characteristics that provide the possibility of uncovering different cyclic aspects in data, 3) developing novel methods for measuring the stability of rhythms, and 4) advancing the machine learning component to incorporate a comprehensive selection of analytic methods that further enhances the capabilities of the framework to be used for predictive modeling of cyclic biobehavior.

5.1 Limitations and Future Work

While the proposed computational framework is easily extendable to various types of data, the current implementation has a few constraints. First, the input sensor signals should be equidistantly sampled for the rhythm modeling methods to work. Second, the input sensor signals need to have significant cyclic patterns. Finally, the length of input sensor signals should be substantially longer than the expected periods to ensure stable modeling. For the current implementation, we also limited our periodic functions to Autocorrelation, Periodogram, and Cosinor. In future work, we hope to build an ensemble system incorporating different types of rhythm detection algorithms and design a voting algorithm for aggregating the outputs of period detection algorithms. For example, the most frequently detected period by various detection algorithms will be treated as the dominant period. We also plan to extend the framework by adding and evaluating novel methods to quantify the collective stability of individual and group rhythms.

For handling missing values, we used nearest-neighbor linear interpolation as one of the fundamental missing data imputation methods. However, we acknowledge that missing data is a daunting issue in sensor data processing, and the strategy for handling missing data requires careful consideration. For example, each sensor stream may have a certain distribution pattern that requires a different handling method. In cases of large continuous missing blocks (e.g., 7 or 10 straight days), interpolation can result in smoothed distributions that do not reflect the actual data and lead to misinterpretations of the built models. In our cases, we set a threshold of 30% to eliminate features with large blocks of continuous missing values and to avoid the above-mentioned problem. The threshold was decided based on our calculation of the length of missing blocks. While this strategy can be useful for many types of data, it may not serve as optimum for all.

Finally, although we presented two cases to demonstrate the capability of the framework in modeling different types of data, more evaluations are needed to verify its generalizability.

6 CONCLUSION

We designed and presented a computational framework for modeling biobehavioral rhythms from mobile and wearable data streams that rigorously process sensor streams, detect periodicity in data, model rhythms from that data and use the cyclic model parameters to predict an outcome. Our evaluation of the framework using two different case studies showed that in addition to detection of rhythmicity, the framework can reliably discover various periods of different lengths in data, extract cyclic biobehavioral characteristics through exhaustive modeling of rhythms for each sensor feature; and provide the ability to use different combinations of sensors and data features to predict an outcome. The machine learning analyses for predicting mental health and readiness demonstrated the ability of our framework to process massive numbers of data streams to build and analyze micro-rhythmic models for each sensor feature and combinations of features and highlighted dominant rhythmic features for prediction of the outcome of interest. The case studies also provided novel findings that were not observed in similar studies. These results show the feasibility of our computational modeling framework for studying different outcomes and extracting new knowledge through modeling biobehavioral rhythms. Further evaluations can verify the generalizability of the framework.

REFERENCES

[1] Saeed Abdullah, Mark Matthews, Elizabeth L. Murnane, Geri Gay, and Tanzeem Choudhury. 2014. Towards circadian computing: "Early to bed and early to rise" makes some of us unhealthy and sleep deprived. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. 673–684.

[2] Saeed Abdullah, Elizabeth L. Murnane, Mark Matthews, and Tanzeem Choudhury. 2017. Circadian computing: Sensing, modeling, and maintaining biological rhythms. In *Mobile Health*. Springer, 35–58.

[3] Saeed Abdullah, Elizabeth L. Murnane, Mark Matthews, Matthew Kay, Julie A. Kientz, Geri Gay, and Tanzeem Choudhury. 2016. Cognitive rhythms: Unobtrusive and continuous sensing of alertness using a mobile phone. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. 178–189.

[4] Aaron T. Beck, Robert A. Steer, Roberta Ball, and William F. Ranieri. 1996. Comparison of Beck depression inventories-IA and-II in psychiatric outpatients. *Journal of Personality Assessment* 67, 3 (1996), 588–597.

[5] Giannina J. Bellone, Santiago A. Plano, Daniel P. Cardinali, Daniel Pérez Chada, Daniel E. Vigo, and Diego A. Golombek. 2016. Comparative analysis of actigraphy performance in healthy young subjects. *Sleep Science* 9, 4 (2016), 272–279.

[6] Peter Bloomfield. 2004. *Fourier Analysis of Time Series: An Introduction*. John Wiley & Sons.

[7] Amy M. Bohnert, Julie Wargo Aikins, and Nicole T. Arola. 2013. Regrouping: Organized activity involvement and social adjustment across the transition to high school. *New Directions for Child and Adolescent Development* 2013, 140 (2013), 57–75.

[8] Michael J. Bradburn, Jonathan J. Deeks, Jesse A. Berlin, and A. Russell Localio. 2007. Much ado about nothing: A comparison of the performance of meta-analytical methods with rare events. *Statistics in Medicine* 26, 1 (2007), 53–77.

[9] Leo Breiman. 2001. Random forests. *Machine Learning* 45, 1 (2001), 5–32.

[10] Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. 2013. API design for machine learning software: Experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*. 108–122.

[11] John Parker Burg. 1972. The relationship between maximum entropy spectra and maximum likelihood spectra. *Geophysics* 37, 2 (1972), 375–376.

[12] Yvonne Chan and Roy P. Walmsley. 1997. Learning and understanding the Kruskal-Wallis one-way analysis-of-variance-by-ranks test for differences among three or more independent groups. *Physical Therapy* 77, 12 (1997), 1755–1761.

[13] Germaine Cornelissen. 2014. Cosinor-based rhythmometry. *Theoretical Biology and Medical Modelling* 11, 1 (2014), 1–24.

[14] Maria J. Costa, Bärbel Finkenstädt, Véronique Roche, Francis Lévi, Peter D. Gould, Julia Foreman, Karen Halliday, Anthony Hall, and David A. Rand. 2013. Inference on periodicity of circadian time series. *Biostatistics* 14, 4 (2013), 792–806.

[15] Pietro Cugini. 1993. Chronobiology: Principles and methods. *Annali–Istituto Superiore Di Sanita* 29 (1993), 483–483.

[16] T. G. Dietterich. 2000. Ensemble methods in machine learning. *Multiple Classifier Systems: First International Workshop, MCS 2000, Lecture Notes in Computer Science*, 1–15.

[17] Yipeng Ding and Jingtian Tang. 2014. Micro-Doppler trajectory estimation of pedestrians using a continuous-wave radar. *IEEE Transactions on Geoscience and Remote Sensing* 52, 9 (2014), 5807–5819.

[18] Afsaneh Doryab, Prerna Chikarsel, Xinwen Liu, and Anind Day. 2018. Extraction of behavioral features from smartphone and wearable data. (12 2018).

[19] Afsaneh Doryab, Anind K. Dey, Grace Kao, and Carissa Low. 2019. Modeling biobehavioral rhythms with passive sensing in the wild: A case study to predict readmission risk after pancreatic surgery. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 1, Article 8 (March 2019), 21 pages. https://doi.org/10.1145/3314395

[20] Afsaneh Doryab, Daniella K. Villalba, Prerna Chikersal, Janine M. Dutcher, Michael Tumminia, Xinwen Liu, Sheldon Cohen, Kasey Creswell, Jennifer Mankoff, John D. Creswell, et al. 2019. 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 and uHealth* 7, 7 (2019), e13209.

[21] Harold B. Dowse. 2009. Chapter 6 analyses for physiological and behavioral rhythmicity. In *Computer Methods, Part A*. Methods in Enzymology, Vol. 454. Academic Press, 141–174. https://doi.org/10.1016/S0076-6879(08)03806-8

[22] David J. A. Dozois, Keith S. Dobson, and Jamie L. Ahnberg. 1998. A psychometric evaluation of the Beck depression inventory–II. *Psychological Assessment* 10, 2 (1998), 83.

[23] Kieron D. Edwards, Ozgur E. Akman, Kirsten Knox, Peter J. Lumsden, Adrian W. Thomson, Paul E. Brown, Alexandra Pokhilko, Laszlo Kozma-Bognar, Ferenc Nagy, David A. Rand, et al. 2010. Quantitative analysis of regulatory flexibility under changing environmental conditions. *Molecular Systems Biology* 6, 1 (2010).

[24] J. T. Enright. 1965. The search for rhythmicity in biological time-series. *Journal of Theoretical Biology* 8, 3 (1965), 426–468.

[25] J. R. Fernández, R. C. Hermida, and A. Mojón. 2009. Chronobiological analysis techniques. Application to blood pressure. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 367, 1887 (2009), 431–445.

[26] Denzil Ferreira, Vassilis Kostakos, and Anind Dey. 2015. AWARE: Mobile context instrumentation framework. *Frontiers in ICT* 2 (05 2015). https://doi.org/10.3389/fict.2015.00006

[27] Jerome Friedman. 2001. Greedy function approximation: A gradient boosting machine. *Annals of Statistics* 29 (10 2001), 1189–1232. https://doi.org/10.2307/2699986

[28] John Gale, Heather Cox, Jingyi Qian, Gene Block, Christopher Colwell, and Aleksey Matveyenko. 2011. Disruption of circadian rhythms accelerates development of diabetes through pancreatic beta-cell loss and dysfunction. *Journal of Biological Rhythms* 26 (10 2011), 423–33. https://doi.org/10.1177/0748730411416341

[29] Quentin Geissmann, Luis Garcia Rodriguez, Esteban J. Beckwith, and Giorgio F. Gilestro. 2019. Rethomics: An R framework to analyse high-throughput behavioural data. *PloS One* 14, 1 (2019).

[30] Anne Germain and David Kupfer. 2008. Circadian rhythm disturbances in depression. *Human Psychopharmacology* 23 (10 2008), 571–85. https://doi.org/10.1002/hup.964

[31] M. Gleicher, T. Landesberger von Antburg, and I. Viola. [n.d.]. ARGUS: An interactive visual analytics framework for the discovery of disruptions in bio-behavioral rhythms. ([n. d.]).

[32] Franz Halberg. 1969. Chronobiology. *Annual Review of Physiology* 31, 1 (1969), 675–726.

[33] Johnni Hansen. 2017. Night shift work and risk of breast cancer. *Current Environmental Health Reports* 4, 3 (2017), 325–339.

[34] Elizabeth Klerman, Andrew Phillips, and Matt Bianchi. 2016. Statistics for sleep and biological rhythms research: Longitudinal analysis of biological rhythms data. *Journal of Biological Rhythms* 32 (10 2016). https://doi.org/10.1177/0748730416670051

[35] Fulton Koehler, F. K. Okano, Lila R. Elveback, Franz Halberg, and John J. Bittner. 1956. Periodograms for the study of physiologic daily periodicity in mice and in man; with a procedural outline and some tables for their computation. *Experimental Medicine and Surgery* 14 1 (1956), 5–30.

[36] Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. 2004. Estimating mutual information. *Phys. Rev. E* 69 (Jun 2004), 066138. Issue 6. https://doi.org/10.1103/PhysRevE.69.066138

[37] Gloria Kuhn. 2001. Circadian rhythm, shift work, and emergency medicine. *Annals of Emergency Medicine* 37, 1 (2001), 88–98.

[38] Hyun-Ah Lee, Heon-Jeong Lee, Joung-Ho Moon, Taek Lee, Min-Gwan Kim, Hoh In, Chul-Hyun Cho, and Leen Kim. 2017. Comparison of wearable activity tracker with actigraphy for sleep evaluation and circadian rest-activity rhythm measurement in healthy young adults. *Psychiatry Investigation* 14, 2 (2017), 179.

[39] Cathy Lee Gierke and Germaine Cornelissen. 2016. Chronomics analysis toolkit (CATkit). *Biological Rhythm Research* 47, 2 (2016), 163–181.

[40] Nicholas R. Lomb. 1976. Least-squares frequency analysis of unequally spaced data. *Astrophysics and Space Science* 39, 2 (1976), 447–462.

[41] Anmol Madan, Manuel Cebrián, David Lazer, and Alex Pentland. 2010. Social sensing for epidemiological behavior change. *UbiComp'10 - Proceedings of the 2010 ACM Conference on Ubiquitous Computing*, 291–300. https://doi.org/10.1145/1864349.1864394

[42] Janna Mantua, Nickolas Gravel, and Rebecca Spencer. 2016. Reliability of sleep measures from four personal health monitoring devices compared to research-based actigraphy and polysomnography. *Sensors* 16, 5 (2016), 646.

[43] Nicolai Meinshausen and Peter Bühlmann. 2010. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 72, 4 (2010), 417–473. https://doi.org/10.1111/j.1467-9868.2010.00740.x arXiv:https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-9868.2010.00740.x

[44] Scott Menard. 2002. *Applied Logistic Regression Analysis*. Vol. 106. Sage.

[45] Jun-Ki Min, Afsaneh Doryab, Jason Wiese, Shahriyar Amini, John Zimmerman, and Jason I. Hong. 2014. Toss'n'turn: Smartphone as sleep and sleep quality detector. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 477–486.

[46] Marie-Christine Mormont, James Waterhouse, Pascal Bleuzen, Sylvie Giacchetti, Alain Jami, André Bogdan, Joseph Lellouch, Jean-Louis Misset, Yvan Touitou, and Francis Lévi. 2000. Marked 24-h rest/activity rhythms are associated with better quality of life, better response, and longer survival in patients with metastatic colorectal cancer and good performance status. *Clinical Cancer Research* 6, 8 (2000), 3038–3045.

[47] Elizabeth L. Murnane, Saeed Abdullah, Mark Matthews, Tanzeem Choudhury, and Geri Gay. 2015. Social (media) jet lag: How usage of social technology can modulate and reflect circadian rhythms. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. 843–854.

[48] Erika Nelson-Wong, Sam Howarth, David A. Winter, and Jack P. Callaghan. 2009. Application of autocorrelation and cross-correlation analyses in human movement and rehabilitation research. *Journal of Orthopaedic & Sports Physical Therapy* 39, 4 (2009), 287–295.

[49] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[50] M. Poyurovsky, R. Nave, R. Epstein, O. Tzischinsky, M. Schneidman, TRE Barnes, A. Weizman, and P. Lavie. 2000. Actigraphic monitoring (actigraphy) of circadian locomotor activity in schizophrenic patients with acute neuroleptic-induced akathisia. *European Neuropsychopharmacology* 10, 3 (2000), 171–176.

[51] Jonathon Pye, Andrew J. K. Phillips, Sean W. Cain, Maryam Montazerolghaem, Loren Mowszowski, Shantel Duffy, Ian B. Hickie, and Sharon L. Naismith. 2021. Irregular sleep-wake patterns in older adults with current or remitted depression. *Journal of Affective Disorders* 281 (2021), 431–437.

[52] Roberto Refinetti, Germaine Cornélissen, and Franz Halberg. 2007. Procedures for numerical analysis of circadian rhythms. *Biological Rhythm Research* 38, 4 (2007), 275–325.

[53] Roberto Refinetti and Michael Menaker. 1992. The circadian rhythm of body temperature. *Physiology & Behavior* 51, 3 (1992), 613–637.

[54] Alain Reinberg and Israel Ashkenazi. 2003. Concepts in human biological rhythms. *Dialogues in Clinical Neuroscience* 5, 4 (2003), 327.

[55] Daniel W. Russell. 1996. UCLA loneliness scale (Version 3): Reliability, validity, and factor structure. *Journal of Personality Assessment* 66, 1 (1996), 20–40.

[56] Sohrab Saeb, Mi Zhang, Christopher Karr, Stephen Schueller, Marya Corden, Konrad Kording, and David Mohr. 2015. Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: An exploratory study. *Journal of Medical Internet Research* 17 (07 2015). https://doi.org/10.2196/jmir.4273

[57] Jeffrey Scargle. 1989. Studies in astronomical time series analysis. III. Fourier transforms, autocorrelation functions, and cross-correlation functions of unevenly spaced data. *Astrophysical Journal* 343 (08 1989). https://doi.org/10.1086/167757

[58] Arthur Schuster. 1898. On the investigation of hidden periodicities with application to a supposed 26 day period of meteorological phenomena. *Terrestrial Magnetism* 3, 1 (1898), 13–41. https://doi.org/10.1029/TM003i001p00013 arXiv:https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/TM003i001p00013

[59] Phillip G. Sokolove and Wayne N. Bushell. 1978. The chi square periodogram: Its utility for analysis of circadian rhythms. *Journal of Theoretical Biology* 72, 1 (1978), 131–160.

[60] Martin Straume, Susan G. Frasier-Cadoret, and Michael L. Johnson. 2002. Least-squares analysis of fluorescence data. In *Topics in Fluorescence Spectroscopy*. Springer, 177–240.

[61] Gunilla Brun Sundblad, Anna Jansson, Tönu Saartok, Per Renström, and Lars-Magnus Engström. 2008. Self-rated pain and perceived health in relation to stress and physical activity among school-students: A 3-year follow-up. *Pain* 136, 3 (2008), 239–249.

[62] John M. Taub. 1978. Behavioral and psychophysiological correlates of irregularity in chronic sleep routines. *Biological Psychology* 7, 1 (1978), 37–53. https://doi.org/10.1016/0301-0511(78)90041-8

[63] Rui Wang, Fanglin Chen, Zhenyu Chen, Tianxing Li, Gabriella Harari, Stefanie Tignor, Xia Zhou, Dror Ben-Zeev, and Andrew Campbell. 2014. StudentLife: Assessing mental health, academic performance and behavioral trends of college students using smartphones. *UbiComp 2014 - Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. https://doi.org/10.1145/2632048.2632054

[64] Gregory William Yeutter. 2016. *Determination of Circadian Rhythms in Consumer-Grade Actigraphy Devices*. Drexel University.

[65] Ping Zhang. 1993. Model selection via multifold cross validation. *The Annals of Statistics* (1993), 299–313.

[66] Tomasz Zielinski, Anne M. Moore, Eilidh Troup, Karen J. Halliday, and Andrew J. Millar. 2014. Strengths and limitations of period estimation methods for circadian data. *PloS One* 9, 5 (2014).

Received November 2020; revised December 2021; accepted December 2021