Leveraging Routine Behavior and Contextually-Filtered Features for Depression Detection among College Students

XUHAI XU, University of Washington
PRERNA CHIKERSAL, AFSANEH DORYAB, DANIELLA K. VILLALBA, and JANINE M. DUTCHER, Carnegie Mellon University
MICHAEL J. TUMMINIA, University of Pittsburgh
TIM ALTHOFF, University of Washington
SHELDON COHEN, KASEY G. CRESWELL, and J. DAVID CRESWELL, Carnegie Mellon University
JENNIFER MANKOFF and ANIND K. DEY, University of Washington

The rate of depression in college students is rising, which is known to increase suicide risk, lower academic performance and double the likelihood of dropping out of school. Existing work on finding relationships between passively sensed behavior and depression, as well as detecting depression, mainly derives relevant *unimodal features* from a single sensor. However, co-occurrence of values in multiple sensors may provide better features, because such features can describe behavior *in context*. We present a new method to extract *contextually filtered* features from passively collected, time-series mobile data *via* association rule mining. After calculating traditional unimodal features from the data, we extract rules that relate unimodal features to each other using association rule mining. We extract rules from each class separately (*e.g.,* depression vs. non-depression). We introduce a new metric to select a subset of rules that distinguish between the two classes. From these rules, which capture the relationship between multiple unimodal features, we automatically extract *contextually filtered features*. These features are then fed into a traditional machine learning pipeline to detect the class of interest (in our case, depression), defined by whether a student has a high BDI-II score at the end of the semester. The behavior rules generated by our methods are highly interpretable representations of differences between classes. Our best model uses contextually-filtered features to significantly outperform a standard model that uses only unimodal features, by an average of 9.7% across a variety of metrics. We further verified the generalizability of our approach on a second dataset, and achieved very similar results.

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

Additional Key Words and Phrases: Behavior mining, Passive sensing, Depression detection, Association rule mining

ACM Reference Format:
Xuhai Xu, Prerna Chikersal, Afsaneh Doryab, Daniella K. Villalba, Janine M. Dutcher, Michael J. Tumminia, Tim Althoff, Sheldon Cohen, Kasey G. Creswell, J. David Creswell, Jennifer Mankoff, and Anind K. Dey. 2019. Leveraging Routine Behavior and Contextually-Filtered Features for Depression Detection among College Students. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 3, Article 116 (September 2019), 33 pages. https://doi.org/10.1145/3351274

Authors' addresses: Xuhai Xu, xuhaixu@uw.edu, University of Washington, 1410 NE Campus Parkway, Seattle, WA, 98195; Prerna Chikersal; Afsaneh Doryab; Daniella K. Villalba; Janine M. Dutcher, Carnegie Mellon University, Pittsburgh, PA, 15289; Michael J. Tumminia, University of Pittsburgh, Pittsburgh, PA, 15260; Tim Althoff, University of Washington; Sheldon Cohen; Kasey G. Creswell; J. David Creswell, Carnegie Mellon University; Jennifer Mankoff; Anind K. Dey, University of Washington.

2474-9567/2019/9-ART116 $15.00
https://doi.org/10.1145/3351274

1 INTRODUCTION

Major depressive disorder (MDD), also known simply as depression, is a common and consequential health challenge. It is often accompanied by low self-esteem [17], loss of interest in normally enjoyable activities, anxiety [39], low energy, and pain [63, 81]. Recent studies found that MDD affects approximately 216 million people around the world [82]. Lifetime depression rates are higher in the developed world (15%) compared to the developing world (11%) [50]. In 2015, an estimated 6.7% of all U.S. adults and 10.3% of young adults had at least one MDD episode over the past year. Among 6.5% of young adults, it was reported that the depressive episodes resulted in severe impairment [2].

The college experience introduces major stressors that affect young adults in a variety of ways, such as academic participation [42], social interactions, financial conditions, and physical and mental health [46]. Moreover, depression is one of the most common symptoms among people with suicidal propensity [52, 64]. In a report by the American College Health Association in 2018 [3], approximately 13.1% of undergraduate students seriously considered suicide in the past year and 2.0% attempted suicide. Yet many of them do not even realize they are depressed until they begin experiencing severe functional deterioration [22, 46], and even those that do may not seek health treatment [29, 40].

Detecting depression and identifying early signs of depression can mitigate or prevent its negative consequences. However, traditional depressive symptom screening methods mainly rely on periodic self-reports, whose effectiveness is impacted by subjectivity and compliance. There is a growing realization that everyday devices, including mobile phones and wearable devices (*e.g.*, smart bands, smartwatches), that consistently and passively collect behavioral sensor data, can help us to understand the relationship between people's daily behavior and depression [8, 19, 55], complement traditional screening methods and provide the potential for introducing real-time interventions [83, 84]. Over the past few years, different studies have identified significant correlations between mobile sensing data and depression [27, 70, 86]. However, most of the previous work has mainly focused on a single sensor channel such as location [19, 30, 70].

Some previous work has directly combined multiple sensor channels but has treated each sensor channel as a separate feature, which misses the opportunity to capture co-occurrence relationships between sensors (*e.g.*, [76, 84]). Such co-occurrence relationships might be able to boost the performance of machine learning models for depression detection and prediction, and more importantly, might be able to provide better insights for understanding people's depression-related behavior from wearable and mobile sensors. In this paper, we present a new approach to capturing these co-occurrence relationships across sensor channels with the goal of identifying students who experienced depressive symptoms at the end of the semester.

Specifically, we present a new method for generating **contextually filtered features**, which performs better than prior feature selection approaches for accurately detecting depression. Given a data set composed of time-series sensor measurements, unimodal features can be calculated as aggregates over different time periods (*e.g.*, daily or every morning) within a single sensor in that data set. We define **contextually filtered features** to be those features that are calculated as an aggregate (mean or standard deviation) over a *filtered* subset of the time periods in the data, with filtering decisions made based on the values of *context* (as defined by co-occurrence of features), *e.g.*, mean of sleep over the nights only when students are off-campus and have high activity levels. Association Rule Mining [4] (ARM) is used to decide which features to use for contextual filtering and aggregation. For example, suppose ARM identifies the rule "when students are sitting in a classroom, they are also interacting with their phone". The unimodal features for 'sitting in a classroom' and 'interacting with their phone' represent simple averages of how often each happens. The contextual filtered feature calculated from this same rule would be an average of 'interacting with their phone' that only includes days where 'sitting in a classroom' is true.

Contextually filtered features are more likely to be able to capture signals of interest around health behaviors, as demonstrated by the improved performance of our approach. Prior approaches to depression detection have leveraged only unimodal features computed over the entire observation period (*e.g.*, [19, 70]), or have manually created and selected any contextual features (*e.g.*, [87]). In contrast, our method leverages association rule mining [4, 77] to automatically and systematically generate contextual features from extracted behavior rules, which define subsets of time periods relevant to a specific context. The behavior rules generated by our method reveal behavior differences between the depression group (having mild or more severe depressive symptoms, as measured by the Beck Depression Inventory (BDI-II) [11]) and the non-depression group. Prior work in association rule mining directly uses extracted rules for classification using a rules-based classifier, without generating features from them (*e.g.*, [56]). Our approach automatically generates rich contextual features from those rules, which can be used with a wide range of classification algorithms.

Our approach is fully automated. It produces a classifier for detecting whether, at the end of the academic semester, a student will have a BDI-II score greater than 13, indicating mild or more severe depressive symptoms [12, 28]. We demonstrate that it outperforms a standard pipeline with unimodal features calculated on the same data, but not filtered with any specific context. When applied to data collected from 138 students at a US university, the new classifier outperforms the standard classifier by 10.2% on accuracy (81.8% vs. 71.6%) and 9.6% on F1 score (84.3% vs. 74.7%) in detecting post-semester depression. Further, our results are robust. We replicate our method on a second dataset collected at the same university a year later and obtain a classifier with an accuracy of 84.0% and F1 score of 88.1%, which outperforms the baseline by 10.0% and 6.4%, respectively.

The contributions of the paper are as follows:

- We present a new approach to perform rule selection on population subgroups that represent different classes of interest, and a new method to generate contextually filtered behavior features based on the outcomes of traditional rule mining algorithms on mobile sensing data.
- We demonstrate that using our method, the best rules are highly interpretable and can capture students' routine behaviors, as well as behavior pattern differences between a subgroup with depressive symptoms and a subgroup without depressive symptoms. The rules obtained by our algorithm help us to better understand students' life experiences related to depressive symptoms.
- We demonstrate that the best model, trained on the contextual features extracted from these rules, outperforms the baseline (using unimodal features), by an average of 9.7% across a variety of metrics.
- We further verify the generalizability of our method. We first apply the rule selected by our method to an independent, second dataset. On this new dataset, the classifier, trained on features extracted based on the same rules, outperforms the baseline by an average of 5.6%. When we run our entire method on the second dataset (mining rules, selecting rules, selecting contextual features and building a classifier) the resulting classifier outperforms the baseline by an average of 7.1%.

To the best of our knowledge, we are the *first to automatically generate contextually filtered features to improve the classifier accuracy in differentiating between two classes, in our case performing depression detection.* Our results create new opportunities for future research, both for depression and for using rule mining to identify useful behavioral features in domains beyond depression. We first summarize related work on depression detection using mobile sensing, rule mining algorithms, and human routine modeling in Section 2. We introduce our rule mining, selecting and contextually filtered feature extraction algorithm in Section 3. Then, we briefly describe the dataset we apply our approach to, and the implementation of our approach in Section 4. We present the results of applying our approach to this dataset: the selected rules and the contextually filtered features, in Section 5. We discuss some implications from our results, limitations of our work, as well as future directions in Sections 6 and 7. Finally, we summarize our contributions and the implications of our work in Section 8.

2 RELATED WORK

Daily behaviors, such as movement patterns, communication, phone use, physical activity, and sleep, can be sensed by various sensors embedded in smartphones and wearable fitness trackers. Features captured by sensors can be indicative of behavioral symptoms of depression. A rich body of work has demonstrated the feasibility of identifying depressive symptoms through passive sensing of daily behaviors [19, 26, 43, 70, 87]. Knowing how a person, with or without depressive symptoms, enacts their daily routine may provide another perspective to understand and detect depression.

Time series data, such as that used in the work modeling depression, has been modeled using a range of algorithms. Techniques that have been applied to model human behavior include Association Rule Mining (ARM) [61, 71–73], T-patterns [16], topic models [31] and maximum causal entropy [9, 10]. Of these, ARM is particularly well suited to generating contextual features, because it is used to find frequent patterns, correlations, or associations between features. In this section, we review and summarize the related literature on depression detection *via* mobile sensing, association rule mining, particularly on rule selection and classification, as well as other techniques for human routine behavior modeling.

2.1 Depression and Mobile Sensing

Recent successes in using mobile sensing to understand and detect depression have made the topic increasingly interesting for researchers. Initial work in this domain explored the statistical relationship between behavior features from mobile sensors and depression (*e.g.*, depression severity) while more recent advances have demonstrated machine learning models that successfully use mobile sensing features for depression detection.

Katikalapudi et al. [49] ran a month-long study with 216 college students and found a direct relationship between depressive symptoms and Internet use. Saeb et al. [70] collected location and phone usage data from 28 adults over two weeks. Their analysis suggested that location features, such as location variance, location entropy and circadian movement (regularity in 24-hour rhythm), are related to depressive symptoms. They also found phone usage features, usage duration and usage frequency to be significantly correlated to depression scores. Wang et al. [86] conducted the StudentLife study at Dartmouth, involving 48 students over a 10-week term. Their analysis revealed significant correlations between depression scores and conversation duration, sleep duration, and frequency and number of Bluetooth encounters. Further analysis of this dataset also found significant relationships between changes in depression scores and features such as sleep duration, speech duration, and mobility [13].

Beyond the work in identifying relationships between mobile sensor data and depression scores, past work has used those relationships to develop machine learning models to detect depression. Saeb et al. [70] employed location features extracted from 28 adults' data over a two-week period. The model trained on their features achieved a leave-one-out average accuracy of 86.5% for distinguishing between participants with and without depressive symptoms. Similarly, Canzian and Musolesi [19] collected data from a group of 28 participants, and trained models on an individual level with location features to detect periods when users experienced depression. Their model achieved 0.71 sensitivity and 0.87 specificity scores. Farhan et al. [30] detected biweekly depression college students using a dataset involving 79 college students over eight months. Their features were extracted from location data and their model achieved an F1 score of 0.82. In addition to using the single sensor channel of location data to detect depression, some work has also combined data from multiple sensors to do this. Wahle et al. [84] trained models on 36 participants over ten weeks. They used unimodal features from location, physical activity, phone usage, calls, messages, and WiFi scans, and achieved an accuracy of 61.5%. Wang et al. [87] detected depression on a weekly basis using features from smartphone and wearable data collected from 68 college students over two nine-week terms. They achieved 81.5% recall and 69.1% precision.

A rich body of literature has suggested the importance of contextual information (*e.g.*, [1, 25]) for depression detection [18, 83]. Nevertheless, most of the previous work involving machine learning techniques either uses data from a single sensor (*e.g.*, location) or directly combines unimodal features from multiple sensors data into a larger input vector to train models. Very little work has investigated the use of multi-channel (or sensor) co-occurrence information to leverage context at the feature level effectively. Wang et al. [87] use some hand-crafted multi-channel features based on DSM-5 (Diagnostic and Statistical Manual of Mental Disorders) depressive symptoms [7], such as phone usage in study spaces for the symptom of diminished ability to concentrate, and dwell time, phone usage, and conversation in social spaces for the symptom of diminished interest in activities. Their features were intuitive and easy to interpret. The features in our work are similar in their interpretability. However, their feature extraction process requires expert domain knowledge and a tedious trial-and-error process. To the best of our knowledge, we are the first to propose a computational method that can automatically identify the intrinsic relationships between multiple sensors and extract **contextually filtered features** (as we defined in Section 1) for depression detection.

2.2 Association Rule Mining and Classification

Association Rule Mining (ARM) is a data mining method that can find frequent patterns, correlations, associations, or causal structures from datasets. It has been used in a range of domains, from helping to discover sales correlations in transaction datasets [65] to identifying disease correlations in medical datasets [6].

ARM outputs frequent co-occurrence patterns expressed as association rules [4]. Each association rule is of the form $[X \rightarrow Y]$, where X and Y are co-occurring sets of context features, and the association rule indicates that the context features in Y are likely to occur whenever the features in X are observed. For rule $[X \rightarrow Y]$, two parameters are defined [4]:

- **Support**: Support represents the fraction of times the context set $\{X, Y\}$ occurs in the dataset, *i.e.*, the joint probability $sup = P(X, Y)$.
- **Confidence**: Confidence represents the proportion of times Y co-occurs whenever X occurs, *i.e.*, the conditional probability $conf = P(Y|X)$.

A support and confidence threshold, namely sup_{min} and $conf_{min}$, is applied to the rule miner, indicating the minimum support and confidence allowable for each discovered rule.

2.2.1 *Selecting Rules.*

A key drawback of ARM is that it usually generates a large number of rules, so that it becomes difficult to identify useful rules. In the data mining community, several researchers have focused on techniques for selecting the top rules, such as ranking rules based on measures such as confidence [20], selecting a set of rules to fulfill some interestingness criteria such as conciseness, coverage, *etc.* [38, 45, 62] or minimizing the redundancy based on rule redundancy criteria [53, 57, 93, 94]. While high support and confidence may be desirable in traditional data mining [4, 5], a high support value in mobile context data mining results in rules which have broad preconditions and are less useful because they capture common daily routines rather than things relevant to the detection of a variable of interest [72]. The existing rule selection metrics do not explicitly emphasize the contextual specificity of the rules or focus on the ability of rules to differentiate between two classes, which is important for capturing behavior pattern differences between sub-groups of interest.

2.2.2 *Using Rules for Classification.*

Liu et al. [56] introduced class association rules, a modification to ARM that outputs $[X \rightarrow y]$ where X is the item set and y is the class label. They define *support* and *confidence* at the rule level as follows: Support for a rule is the proportion of instances in the data set that match X; Confidence is the ratio of instances which match X and are labeled y divided by all instances that match X. Rules whose support and confidence are above a pre-determined threshold are selected and used as binary features for a rule-based classifier. Some follow-up work aimed to improve this algorithm's efficiency [90, 95], but did not significantly change the prediction mechanism. Other work has combined class association rules with classifiers such as logistic regression [47], decision trees [68] and support vector machines [51]. However, these methods mainly integrate rules into the model directly and do not post-process the original features based on the rules (*e.g.*, use the rule-based feature vector to train SVM classifier [51]). This can miss intrinsic co-occurrence relationships between the context X and Y. We show in Section 5 that a rule-based classifier does not have as good performance as contextually filtered features for depression detection.

2.3 Modeling Human Routine Behavior

While the ability to use ARM for classification is intriguing, the rules generated by ARM are also intrinsically interesting because of their ability to describe and model human behavior. A clear picture of the many aspects of routine behavior can help researchers to generate theories and models of human behavior [9]. It may also provide a better understanding of depressive symptoms. A variety of data mining and machine learning techniques have been used to model human routines. For example, Brdiczka et al. [16] use T-patterns [58] to automatically find recurrences of events in behavior logs. Farrahi and Gatica-Perez [31] use n-gram topic models [88] to capture user context and actions. Banovic et al. [9] employ the maximum causal entropy algorithm [97] to model routine behaviors and their variance. Pierson et al. [66] propose Cyclic Hidden Markov Models to model cycles in human behavior.

Among these options, ARM is a powerful method for mining contextual data to better understand human behavior. Nath [61] describe the ACE system (Acquisitional Context Engine), which uses ARM to mine co-occurrence patterns amongst context events, and exploits the resulting patterns to speculatively sense user context in an energy-efficient manner. Srinivasan et al. [73] present the MobileMiner system, which runs on a smartphone and can discover frequent co-occurrence patterns indicating which context events frequently occur together. Recently, the authors employed the same algorithm and the RuleSelector system [72] to capture user context and to allow smartphone users to browse, modify, and select action rules from a small set of summarized rules presented to the user in a manner similar to the IFTTT (*If-This-Then-That*) platform.

Previous work has also directly leveraged the outcomes of rule mining algorithms to perform behavior prediction. This approach has worked for simple behavior modeling such as phone application usage or phone charging behavior [72, 73]. Since this method mines behavior of multiple users together, without treating subgroups of interest separately, it can easily overlook the behavior differences between groups, especially when dealing with complex behaviors such as depressive symptoms.

3 RULE MINING, SELECTION, AND MULTIMODAL FEATURE EXTRACTION ALGORITHM

In this section, we introduce our method that can capture behavior differences between two groups of users using mobile behavioral data. We introduce our algorithm to extract contextually filtered features based on rule mining and rule selection for a classification problem. Figure 1 visualizes the overall pipeline for our method. We first extract unimodal features and then use ARM to generate rules, on a per-class basis, using those features. We then use a novel metric to select the top rules. The key idea of our approach is to identify rules that identify important differences between classes of users. The classes can be different by (1) sharing similar contexts but with different behavior in those contexts (see Section 3.2.1), or (2) having contexts that are common in one class and uncommon in the other class (see Section 3.2.2). Based on the top rules, we describe a new, automated approach to obtain contextually filtered features based on the top rules (see Section 3.3).

3.1 Step 1: Rule Mining in Two Classes Separately

We first split the dataset into two groups according to our classification label, namely *grp1* and *grp2*. (*i.e.*, depression group versus non-depression group). We perform ARM on them separately to generate a large rule set in each

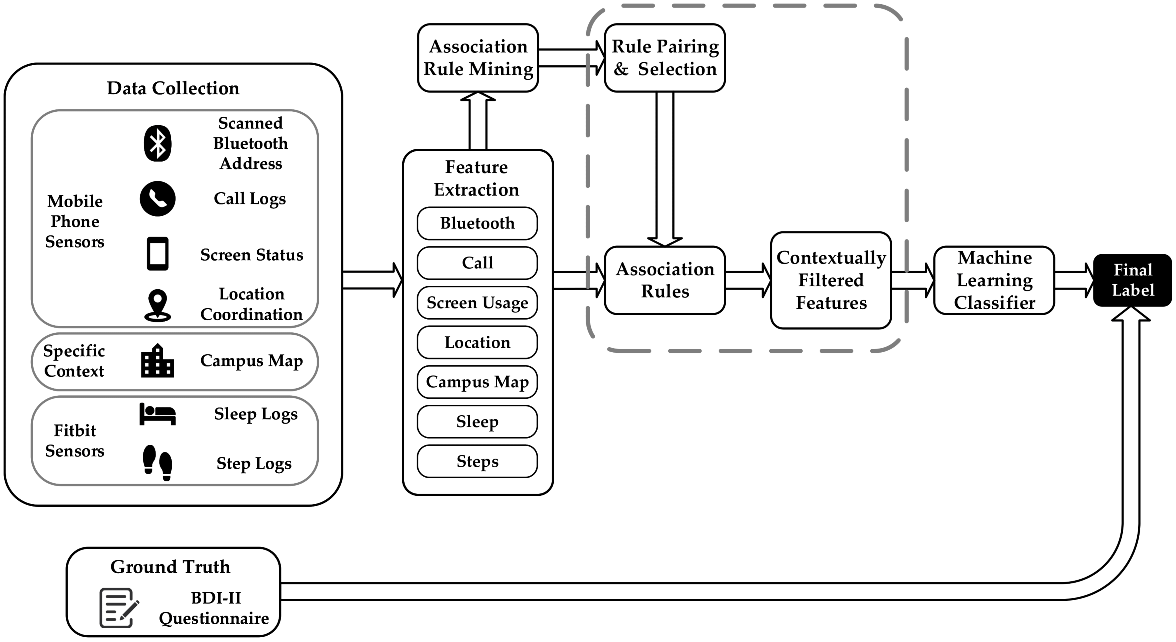


Fig. 1. The high-level pipeline of the integration of rule mining algorithms and machine learning models. The dashed frame highlights the novel contribution of the paper. We designed a new metric to select the top rules from the rule set generated by ARM. We also proposed a new approach to extract contextually filtered features based on the top rules. Finally, we use these features to train classifiers.

group. ARM naturally fits our problem since we obtain multiple features from various sensors at the same time. In a rule $[X \rightarrow Y]$, both X and Y would contain behavioral features. An example rule could be: [X: {*Staying at home, Low activity level*} → Y: {*Being asleep*}] during the night. Next, we devise a novel approach to select the best rules from the two rule sets.

3.2 Step 2: Rule Selection Using a Novel Metric

As described above, our method emphasizes the characteristics of and differences between the two classification groups. Among the two rule sets generated from the two groups described in Section 3.1, there can be rules that are discovered in both groups, capturing common contexts between both groups, that we can use to see how the two groups behave differently; and rules that only appear in one group, that indicate contexts that vary between the groups. To capture the difference, we use two complementary perspectives: one looks at rules that are the same between two groups but with different *sup* and *conf* values, while the other looks at rules that are unique to only one group.

3.2.1 Common Rules in Two Groups. Mobile behavioral data is likely to include many common behavior patterns. Thus, the set of rules present in each group may be very similar. However, even if a rule is present in both groups, *sup* and *conf* can be very different in the groups. We present a set of metrics for characterizing the usefulness of a rule for identifying group membership of an individual.

Contextual Specificity. Rules that are too general are unlikely to discriminate between the two groups. We filter rules that are not very specific. For example, a rule that captures a behavior pattern such as changing locations every weekday morning (from home to work) is less specific than a rule that is more specific, *e.g.*, changing location with a low level of physical activity but a high number of co-locations with other people (the number of Bluetooth encounters). We formalize this in terms of the number of features in X for a rule $[X \longrightarrow Y]$.

$$CtxSpec = |X|$$

Confidence Difference. The confidence of a rule, *i.e.*, the conditional probability, indicates how probable Y is to occur given the context feature set X. For the same X, the difference in confidence directly reflects the different probabilities of Y between two groups. For example, when in working spaces such as offices or libraries (same X), people with depressive symptoms may have more difficulties with concentration [7], thus spending more time interacting with their phones [87]; this could appear in our analysis as higher confidence for the behavior rule R_i $[X\colon \{\textit{Stay at working spaces}\} \longrightarrow Y\colon \{\textit{Phone interaction time}\}]$ for people with depressive symptoms.

The bigger the difference in confidence, the greater the discrepancy in the expression of the rule between the two groups. We formalize this as

$$ConfDiff = |\Delta conf|$$

Condition Discrepancy. The probability of the context set X, *i.e.*, $P(X)$, closely interrelates with the confidence of the rule. We are interested in rules that have different context probability across groups. Continuing the previous example, people with depressive symptoms may spend less time in working or social spaces [7], leading to a different $P(X)$ for those with and without depressive symptoms for rule R_i. Note that $P(X) = \frac{P(X,Y)}{P(Y|X)} = \frac{sup}{conf}$. Thus we formalize this as

$$CondDisc = |\Delta \frac{sup}{conf}|$$

Direction Difference. We would like to find the rules that have *CondDisc* and *ConfDiff* in the same direction. In other words, we are interested in a rule that has both higher $P(X)$ and higher $P(Y|X)$ in one group than the other. We formalize this as

$$DirDiff = \begin{cases} 1 & \text{if } \mathrm{sign}(\Delta conf \cdot \Delta \frac{sup}{conf}) \text{ is positive} \\ 0 & \text{otherwise} \end{cases}$$

Based on these characteristics, we combine the four characteristics into a metric M using Equation 1. The intuition comes from a weighted addition of the logarithm value of the three characteristics. The logarithmic function is monotonically increasing, thus it will not change the relative order when ranking the rules based on the metric.

$$M \;=\; DirDiff \cdot \; CtxSpec^{w_1} \;\cdot\; ConfDiff^{w_2} \;\cdot\; CondDisc^{w_3} \tag{1}$$

DirDiff simply causes features to be dropped if a rule has reverse directions on the *ConfDiff* and *CondDisc* between two groups. The three weight values are used to adjust the relative importance of the remaining three characteristics. We rely on M to rank the rules that are common in both groups and select the top-n rules. We remove redundant rules by the definition: $Rule_1 \; covers \; Rule_2 \iff X_2 \subseteq X_1$ and $Y_2 \subseteq Y_1$. Algorithm 1 (top) presents the procedure.

3.2.2 Unique Rules in One Group. In addition to common rules that are mined from both groups, the two groups can also have unique rules discovered in one group but not the other. These rules reflect the differences of behavior patterns and contexts between the two groups. Selecting the best unique rules in each group can also help identify the distinctions between the two groups. *ConfDiff*, *CondDesc* and *DirDiff* are undefined when a rule is present in only one group. We solve this by setting $\frac{sup}{conf}$ and *conf* to zero when a rule is not present, thus simplifying the calculation of *ConfDiff* and *CondDesc* as shown in Equation 2. *DirDiff* is not used since it always equals 1. We then employ the same metric M as Equation 1 and select the top-n rules.

$$\begin{aligned} ConfDiff &= conf \\ CondDisc &= \frac{sup}{conf} \end{aligned} \tag{2}$$

However, note that the approach above could treat a rule as being unique to one group if the threshold values filtered the rule out in the other group. For instance, take a rule r that occurs in both groups with $sup_1 = 0.101$ and $sup_2 = 0.099$, which are very close. The rule would not appear in *grp2* when the threshold $sup_{min} = 0.100$ is applied, while it would appear in *grp1*.

Thus we need to filter out these rules whose *sup* and *conf* are close to the threshold. We set the minimal distance to be 99th percentiles of the $|\Delta sup|$ and $|\Delta conf|$ we observe from Section 3.2.1 which shows their distributions, and filter out rules that are close to this threshold, as shown in lines 15-18 of Algorithm 1. Similar to the common rules, we rank the resulting unique rules using metric M, and remove any redundant rules.

Overall, we obtain T_{common} from Section 3.2.1 and T_{unique} from Section 3.2.2. The final set of top rules is calculated on line 28 as $T_{top} = T_{common} \cup T_{unique}$. Next, we describe a new approach to extract contextually filtered features from the top rule set.

3.3 Step 3: Contextually Filtered Feature Creation

Once a top rule set T_{top} has been selected using the algorithms described in Section 3.2, they can be used to generate contextually filtered features. These features in turn can be fed into a machine learning model to train a classifier.

For each rule $[X \longrightarrow Y]$, we use X as the "selector" (or filter) to select the days over which to aggregate (the days that fulfill the context feature sets, *i.e.*, the elements of $[X]$). For each element of $[Y]$, we calculate the mean and standard deviation using data from all of the filtered days. Consider as an example, the rule [X: {*Being at sport spaces, High activity level* } $\longrightarrow$ Y: {*Long phone call duration*}]. We select all time periods, or epochs (described in Section 4.2), E_p for person p that fulfill the context {*Being at sport spaces, High activity level* }. Then, we calculate the average and standard deviation of the features in Y only for the selected epochs. Thus in this example, the new contextually filtered features for the person p are the mean and standard deviation of *Duration of out-going call*, for all epochs in which the person spent a long time in sport spaces and had a high level of physical activity.

Algorithm 2 presents the feature extraction procedure. Note that one rule can have multiple features in set Y, thus the number of features generated can be greater than the number of rules. There can also be duplicate features y in Y for different rules. However, as the context (X) of these rules are different, the contextually filtered feature calculated from the same y in different rules is expected to have different mean and standard deviation values.

The final feature set for each person can be used for training classifiers to identify group membership of a person. In the rest of the paper, we describe how we apply our approach for rule mining, rule selection and contextually filtered feature extraction on our dataset, with a focus on depression detection among undergraduate students.

4 TWO YEAR DATA COLLECTION WITH FIRST AND SECOND YEAR STUDENTS

In this section, we briefly describe the depression dataset that we use to demonstrate the effectiveness of our rule selection and contextually filtered feature extraction algorithms. Our data collection was inspired by and modeled after the work of Wang et al. [86]. We describe our specific approach in more depth in the following sections.

```
Data: $grp1$, $grp2$, mining thresholds $sup_{min}$ and $conf_{min}$
1  $R_1$ = ARM($grp1$, $sup_{min}$, $conf_{min}$);
2  $R_2$ = ARM($grp2$, $sup_{min}$, $conf_{min}$);

   // Select Common Rules, see Section 3.2.1
3  $R_{common} = (R_1 \cap R_2)$;
4  for each rule $r$ in $R_{common}$ do
5      $CtxSpec = |X|$;
6      $ConfDiff = |\Delta conf|$;
7      $CondDisc = |\Delta \frac{sup}{conf}|$ ;
8      $DirDiff$ = sign($\Delta conf \cdot \Delta \frac{sup}{conf}$) $> 0$;
9      $M_{common}[r] = DirDiff \cdot CtxSpec^{w_1} \cdot ConfDiff^{w_2} \cdot CondDisc^{w_3}$
10 end
11 $sort(M_{common})$ ;                                                          // Sort by score
12 $T_{common} = M_{common}[0] ... M_{common}[n-1]$ ;                            // Select top $n$ rules
13 $T_{common} = T_{common} \setminus \{r \in T_{common} \mid \exists\, r^* \in T_{common}, r^* \neq r, X_r \subseteq X_{r^*} Y_r \subseteq Y_{r^*}\}$ ;   // Remove redundancy

   // Select Unique Rules, see Section 3.2.2
14 $R_{unique} = (R_1 \cup R_2) \setminus (R_1 \cap R_2)$;
15 $sup_{dis} = 99^{th}$ percentile($|\Delta sup|$ in $R_{common}$);
16 $conf_{dis} = 99^{th}$ percentile($|\Delta conf|$ in $R_{common}$);
17 $R_{unique} = R_{unique} \setminus \{r \in R_{unique} \mid r_{sup} < sup_{min} + sup_{dis}\}$ ;     // Remove rules close to threshold
18 $R_{unique} = R_{unique} \setminus \{r \in R_{unique} \mid r_{conf} < conf_{min} + conf_{dis}\}$;
19 for each rule $r$ in $R_{unique}$ do
20     $CtxSpec = |X|$;
21     $ConfDiff = conf$;
22     $CondDisc = \frac{sup}{conf}$ ;
23     $M_{unique}[r] = CtxSpec^{w_1} \cdot ConfDiff^{w_2} \cdot CondDisc^{w_3}$
24 end
25 $sort(M_{unique})$;
26 $T_{unique} = M_{unique}[0] ... M_{unique}[n-1]$;
27 $T_{unique} = T_{unique} \setminus \{r \in T_{unique} \mid \exists\, r^* \in T_{unique}, r^* \neq r, X_r \subseteq X_{r^*} Y_r \subseteq Y_{r^*}\}$;

   // Merge The Rules
28 $T_{top} = T_{common} \cup T_{unique}$;
```

Algorithm 1: Given a data set split into two groups based on a label (such as depressive symptoms/no depressive symptoms), select the best rules that are present in both groups (lines 3-13), and unique to one group (lines 14-27).

```
Data: person set $P$, top rule set $T_{top}$
for each person $p \in P$ do
    Let $E_p$ be all epochs involving person $p$;
    Let $E_{p.f}$ be all features in $E_p$, start empty;
    for each rule $r \in T_{top}$ do
        $E_r$ = all epochs $e \in E_p$ where $X$ is fulfilled;
        for each unimodal feature $y \in Y$ do
            $r_{mean(y)} = mean(y, E_r)$ ;
            $r_{std(y)} = std(y, E_r)$;
            $E_{p.f} = E_{p.f} \cup \{r_{mean(y)}, r_{std(y)}\}$;
        end
    end
end
```

Algorithm 2: Extracting contextually filtered features from the top rule set.

4.1 Data Collection

During Phase I of the study, we recruited 188 first-year undergraduate students at a Carnegie-classified R-1 university in the United States *via* emails and Facebook posts. Students were invited to the lab to provide informed consent, download a mobile application to track sensor data from their smartphones, and receive a Fitbit Flex 2 to track their steps and sleep. They were asked to keep the application installed, and wear the Fitbit tracker over one semester (106 days). At the beginning and the end of the semester, they were required to answer a depression assessment questionnaire to evaluate their depressive symptoms. For their participation, the participants were allowed to keep the Fitbit Flex 2 and received up to $205, based on their compliance.

We recruited another group of 267 undergraduate students in Phase II of our study, one year later. 85 participants were return participants from Phase I, 33 participants were new second-year participants, and 149 participants in Phase II were new first-year participants. We used a similar data collection procedure as used in Phase I. We use this second dataset to verify the generalizability of our method.

Table 1. Information of the two studies after removing students who dropped out or were missing a significant amount of data. Students with a post-semester BDI-II score greater than 13 were in the depression group, in accord with the interpretation of the BDI-II [12]. Note that owing to some subtle distinctions of the study goals, we did not collect pre-semester BDI-II scores in Phase II.

Study	Days	Overall	Dropped out	Removed	Dataset	Pre-semester BDI-II		Post-semester BDI-II	
		Number	Number	Number	Size	Non-dep Grp	Dep Grp	Non-dep Grp	Dep Grp
Phase I	106	188	28	22	138	114	24	81	57 (41.3%)
Phase II	113	267	31	24	212	-	-	136	76 (35.8%)

Table 1 summarizes both phases of data collection, which was IRB-approved. During Phase I data collection, 28 students dropped out of the study owing to various personal reason. During data cleaning, we removed 22 further students who we were missing a significant amount of data. 138 students out of 188 recruited were used for analysis. Among the 267 students in the Phase II dataset, 31 dropped out and 24 were removed due to missing data, leaving 212 students.

4.1.1 Ground Truth Data Collection. We employed the Beck Depression Inventory-II (BDI-II) [12], a widely used psychometric test for depressive symptoms severity measurement, to obtain ground truth. The BDI-II has strong psychometric properties [12, 37] and is the most widely used self-report measure of the presence and severity of depressive symptoms in non-clinical samples and clinical trials of depression [37]. Self-reported depression on the BDI-II has been shown to discriminate patients who were diagnosed with mild, moderate, and severe major depressive episodes as defined by the DSM-5 [74]. Importantly for our purposes, many studies have provided validation information and normative data about the BDI-II in large college student samples (*e.g.*, [23, 75, 91, 92]). The questionnaire contains 21 questions, with each answer being scored on a scale of 0-3. 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) [28]. We labeled the students whose post-semester BDI-II score was higher or equal to 14 as having depressive symptoms.

Participants answered the BDI-II questionnaire at the beginning (week 1) and at the end (week 16) of the semester. Forty-one percent of participants in Phase I, and thirty-six percent in Phase II, had depressive symptoms consistent with mild or stronger depression according to their BDI-II scores. This is similar to national rates for depression among college students: as reported in the ACHA-NCHA II [3], 35.6% of college students experienced depression in the past year and 18.4% were diagnosed with depression.

Note that the BDI-II asks about the severity of depressive symptoms *in the past two weeks*. Thus, we use passively sensed data collected throughout the semester to predict participant status on the BDI-II at the end (week 16) of the study.

4.1.2 Passive Mobile Data Collection. We installed the AWARE framework [32] to collect sensor data unobtrusively from students' smartphones. The application recorded students' nearby Bluetooth addresses, call logs, phone usage (charging activity and screen status), and location. Further, participants were required to wear a Fitbit Flex 2 that recorded their steps and sleep status (leftmost part of Figure 1 shows the sensor type). While Calls and Phone Usage are event-based data, Bluetooth, Location, Sleep, and Steps are sampled, time series data. We sampled Bluetooth and Location coordinates at 1 sample per 10 minutes, Sleep at 1 sample per minute, and Steps at 1 sample per 5 minutes. Data from AWARE was de-identified locally on the phone and automatically transferred over Wi-Fi to our back-end server on a regular basis. Data from Fitbit was collected using the Fitbit API at the end of the study. Participants were required to keep their phone and Fitbit charged and carry/wear them at all times.

4.2 Unimodal Feature Extraction

The sensor streams from which we collected data included sleep (Fitbit), steps (Fitbit), Bluetooth (phone), calls (phone), screen use (phone) and GPS (phone). Since we are interested in aggregating sensor values into features for the entire study, we used the approach describe in Wang et al. [86] to group raw sensor data into epochs that capture behavior at different times of day. Past literature has suggested that people usually have different behavior patterns during different times of the day [21], and between weekdays and weekends [69]. Following Wang et al. [86], we divide the data into four epochs for weekdays (night *i.e.*, 12am-6am, morning *i.e.*, 6am-12pm, afternoon *i.e.*, 12pm-6pm, evening *i.e.*, 6pm-12am) and the same four epochs for weekends, resulting in 8 epochs.

We then aggregate sensor streams on a per-day, per-epoch basis into *daily-epoch features* (such as the number of phone calls on the morning of Tuesday, February 18, 2017). Aggregation is done on a per-day basis producing daily-epoch features. For sampled data, such as Bluetooth data, features naturally divide on epoch boundaries. For event-based data, such as phone calls, we use the start time of an event to determine which epoch it belongs to (*e.g.*, a phone call from 11:55 am to 12:05 pm, would be in the morning epoch).

Most features are aggregated using a mix of mean, maximum, minimum, and standard deviation for sampled data, and count and duration for event-based data, if appropriate. However, some features require additional

Table 2. Sensor data and information aggregated into features.

Sensor	Source	Sampling	Information Being Aggregated into Features	Number of Samples Per-person
Screen	AWARE	event-based	Number of unlocks per minute, total time with interaction, total time unlocked	39843.2 ± 22126.9
Call			Number and duration of in-coming /out-going/missed calls	379.6 ± 275.8
Bluetooth		1 per 10 minutes	Number of unique devices, number of scans of most/least frequent device	24579.0 ± 106960.9
Location			GPS latitude, longitude, altitude	9692.8 ± 4444.2
Sleep	Fitbit	1 per minute	Asleep/restless/awake/unknown duration and onset	34963.6 ± 15630.6
Step		1 per 5 minutes	Number of steps	23390.6 ± 10197.7

pre-processing. For location features, we calculate *location variance* (sum of the variance in latitude and longitude coordinates), *total distance traveled, average/variance of speed* and *circadian movement* (the degree that a person's mobility patterns follow a 24-hour circadian cycle [70]). We also cluster locations within a epoch and globally, to determine the *number of significant places, number of transitions between places, radius of gyration* [19], *percentage of time spent at top-3 frequented clusters/moving/rarely visited locations, length of stay at clusters* and *location entropy*. We analyze the user's location patterns in relation to the college campus map, focusing specifically on Greek houses (which tend to hold social events), residential halls, athletic facilities, green spaces, and academic buildings. For sleep, we add sleep efficiency, and sleep onset. For steps, we group activity into active bouts and sedentary bouts (less than 10 steps in a 5-minute interval). For Bluetooth, we cluster devices into frequently seen groups and count prevalence of each cluster.

Finally, we aggregate daily features into *full features*, within an epoch, using the mean and standard deviation of daily-epoch features. For example, the average number of calls (weekday morning epoch) is calculated as an average over the daily-epoch feature which captures the number of calls made each weekday morning (6am-12pm). We calculate a total of 212 daily-epoch features from the raw sensor streams (summarized in Table 2. This results in a total of 212 x 2 (mean and std deviation), or 424 features per epoch. The pipeline introduced in Section 4.3 is run separately on each epoch, and makes use of full features to calculate rules. It then creates *new* full features that aggregate only some days, as described in Section 3.3.

Each sensor stream results in features that are designed to capture behavior variability related to variables that might be influenced by depressive symptoms [7]. For example, depression can cause sleep disturbances [80] and diminished concentration [24]. The former might impact features such as time in bed, while the latter might impact a metric such as phone use, which could rise with distraction and depression [24].

4.3 Pipeline for Detecting Depression

Our pipeline for depression detection is shown in Figure 2. As described next, the data set is split into two parts (Dataset for rule mining, and Dataset for model training). The first is used to calculate and select features; to which the algorithm described in Section 3 is applied. The results of this algorithm are used to create contextually filtered features that are combined with the unimodal features calculated on the dataset for model training. This second data set is then used to train and test a model using AdaBoost [35] with decision-tree-based component classifiers, with leave-one-out cross-validation. More detail on each of these steps is provided below.

4.3.1 Data Set Preparation. To avoid overfitting, we randomly divided our dataset on a per-person basis into two subsets. We created a RuleGenerateSet of 50 people for extracting rules (20 in the depression group; 30 other),

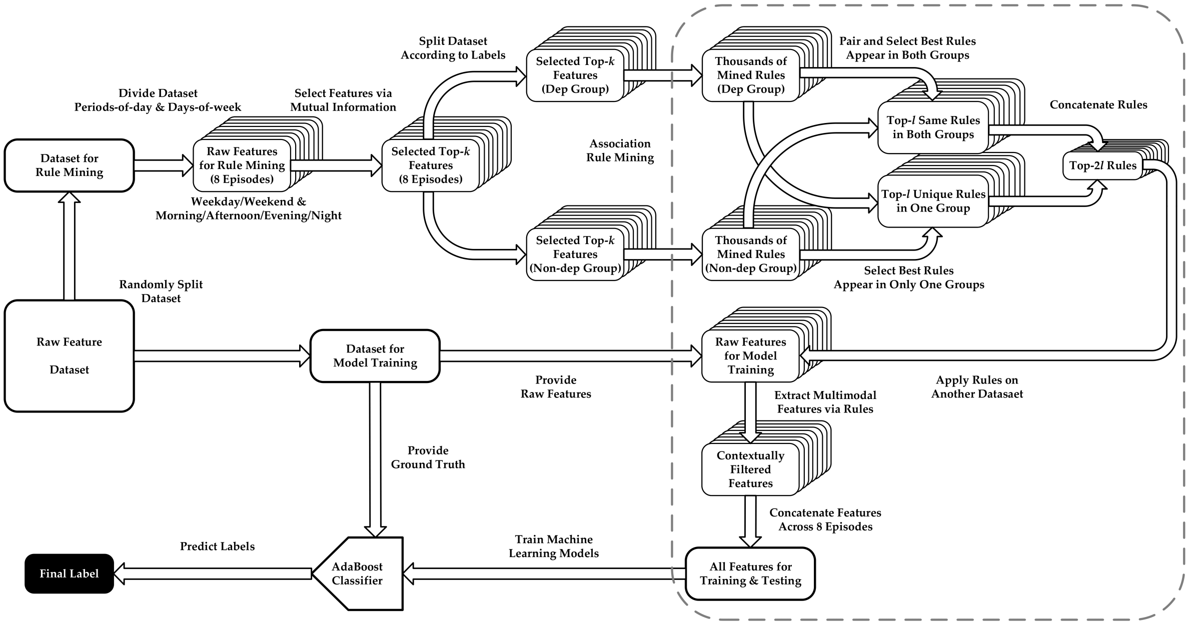


Fig. 2. The detailed pipeline of rule mining, pairing, selecting and models training. The dashed frame highlights the novel procedures in the pipeline.

and a TrainTestSet of 88 people to train and test the machine learning models (37 in the depression group; 51 other). We applied our rule mining and selecting algorithm (Section 3.1 and 3.2) on the RuleGenerateSet and calculated the contextually filtered features (Section 3.3) on the TrainTestSet.

4.3.2 Feature Selection. We selected a subset of the 424 features in each of the eight epochs to reduce computational complexity. We employed mutual information [67] to perform feature selection. We used the method described in [54] to estimate the mutual information gain. Random noise was added to the variables to remove repeated values, thus each calculation batch would have a different ranking order and have a slightly different top feature set. We started with the whole feature set, repeated the calculation and iteratively selected the intersection of the top 50 features until the number of features converged. We ended up with 23 top features on average (Min = 18, Max = 29, 181 in total) among the 8 epochs. We denote these 181 features as the *unimodal feature set*, which we used to train baseline classifiers since this approach is similar to the common practices used in previous literature related to depression detection [84].

We considered the top features in each epoch group to identify the daily-epoch features to be used for rule mining. Specifically, a daily-epoch feature would be selected as long as either the mean or standard deviation of the feature is in the top feature set. We obtained an average of 20 (Min=15, Max=29) top daily-epoch features in each epoch. We used these features for rule mining in Section 4.3.

4.3.3 Feature Preparation before Rule Mining. ARM is typically applied on symbolic or categorical data. We therefore recoded each of the selected features, for rule selection using ARM only, into the three categories: low, moderate, and high, using a binning method. Each category contained 33.3% of each feature, which means the two cut-off thresholds were 33.3 and 66.6 quantiles of the data. Note that since each individual has different behavior patterns, we discretized the data within each individual rather than across individuals. Ideally, each day would contain all of the selected top daily-epoch features. However, sometimes not all features were available due to missing data arising from issues with low smartphone battery, data transfer from the phone to the server, or users not giving permission for certain data to be collected. In each epoch, we filtered out the days where more than half of the features were missing before rule mining.

4.3.4 Rule Mining and Selecting. Once we obtained the discretized top daily-epoch features, we fed them into our pipeline. We employed the tools provided by [33] to mine rules 16 separate times: Once in each of the 8 epochs for each class of users (depressive symptoms and no depressive symptoms). Some epochs (*e.g.*, weekday morning group) would generate over one million rules if their threshold were as low as other epochs. Therefore, we set sup_{min} and $conf_{min}$ in each group separately (0.07-0.19) to control the number of generated rules. We found approximately 16,000 rules (Min = 4,500, Max = 26,000) among the groups. For each epoch, we used Algorithm 1 (top) to select the best common rules, and Algorithm 1 (bottom) to select the best unique rules, for the two classes of participants. Note that we used grid search for Equation 1, ranging from 0.0 to 2.0 with 0.5 as the interval, to set the best weights (w_1, w_2, w_3) which were $(1.0, 1.5, 0.5)$. We used the F1 score in the RuleGenerateSet as the metric for selection (using the same procedure in Section 4.3.5). We obtained an average of 13 rules (Min = 6, Max = 19 rules) per epoch, 105 in total.

4.3.5 Feature Extraction and Model Training. After we obtained the rules, we turned to the TrainTestSet and used Algorithm 2 to extract an average of 17 contextually filtered features (Min = 8, Max = 23) per epoch, 137 in total. Note that one rule $X \rightarrow Y$ can have multiple features y in Y, thus the number of contextually filtered features generated can be greater than the number of rules. We aggregated each y in Y, for each individual, using mean and standard deviation, over daily-epochs that matched X. We added 274 additional features to the unimodal features already available for each participant (137×2). From this, model training can commence.

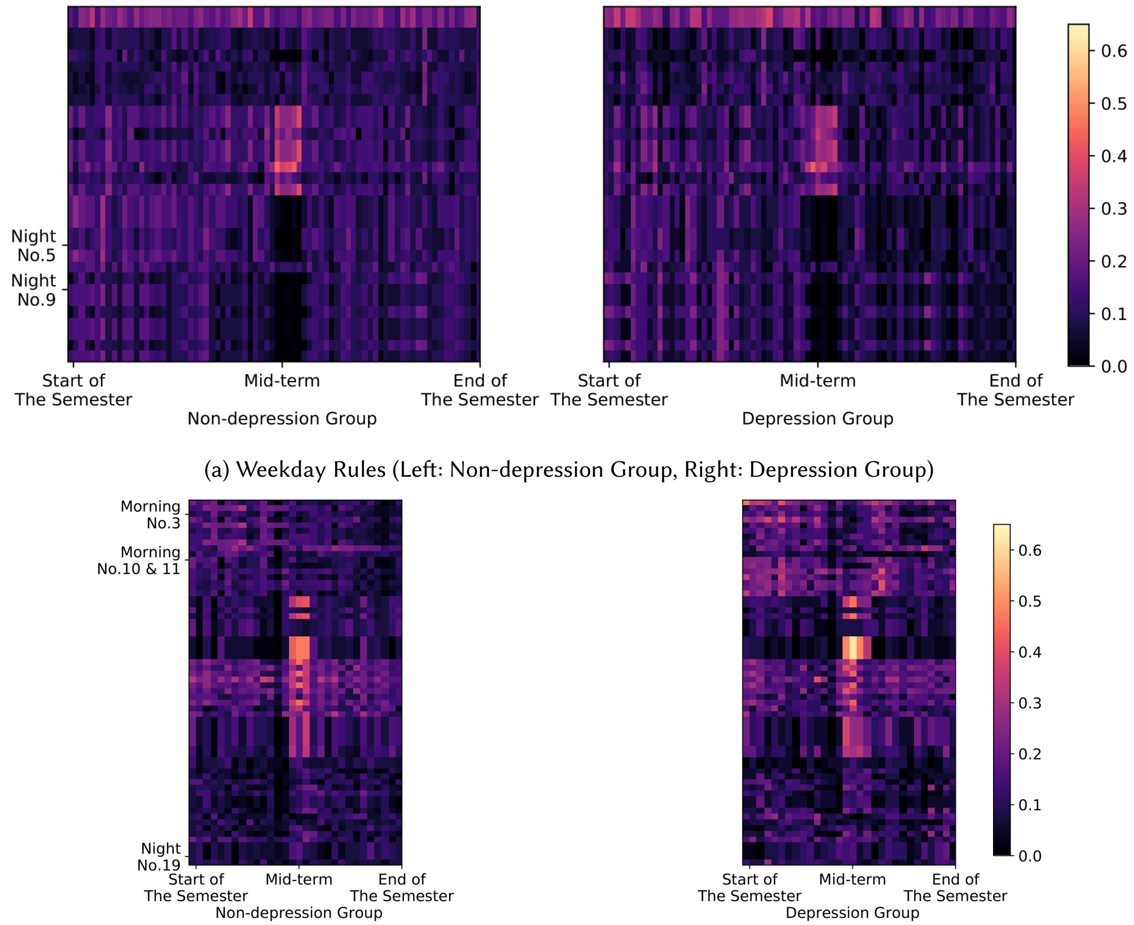
5 VALIDATION OF ALGORITHM

In this section, we verify our methods from several perspectives. We first show in Section 5.1 that the top rules can capture the behavior differences between the student group with depressive symptoms and the student group without depressive symptoms. These results indicate that the rules from our method have good interpretability and can help better understand students' life experiences related to depressive symptoms. Then, in Section 5.2, we demonstrate that the best classifier trained on our contextually filtered features can achieve an average of 9.7% performance increase over the baseline model trained on unimodal features. We further verify the generalizability of our method in Section 5.3. Our rules mined in Section 5.2.1 can be directly applied to a separate dataset to extract contextually-filtered features and train classifiers on that dataset, and the resulting model outperforms the baseline model on the same dataset by 5.6%. We also re-execute our pipeline on the separate dataset, and our best model has an average increase of 7.1% over the equivalent baseline. These results verify the effectiveness and generalizability of our method.

5.1 Rules Can Capture Routines Behaviors and Behavior Pattern Differences between Groups

Our method described in Section 3 aims to find rules that can distinguish between classes of participants. We show that the top rules discovered in our dataset are able to capture students' behavior patterns as well as the behavior differences between students who report depressive symptoms on the BDI-II and those who do not.

Figure 3 visualizes heatmaps that represent how many students' behavior was captured by each of the 105 rules throughout the study period. From both heatmaps of weekdays and weekends, we observed abnormal color patterns during the middle of the study. The academic calendar of the university showed that this period was when midterm examinations took place, followed by a spring break. Students would usually have a stressful period to prepare for examinations, and then have a brief relaxing period. As a result, during the midterm and



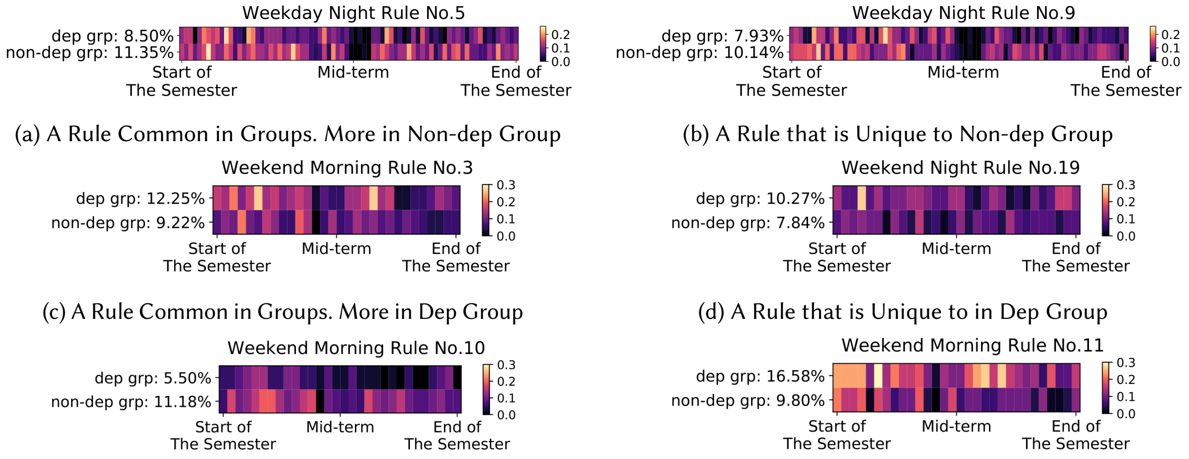
(a) Weekday Rules (Left: Non-depression Group, Right: Depression Group)

(b) Weekend Rules (Left: Students with no depressive symptoms, Right: depressive symptoms)

Fig. 3. Heatmaps of prevalence of the top 105 rules among students with and without depressive symptoms, for weekends and weekdays. X axis is day of semester, Y axis shows rules aligned from morning to night epochs. Color indicates the proportion of students in a class that fulfill a particular rule. The brighter the color, the larger proportion of students having the pattern. The abnormal vertical color patterns in the middle of both figures correspond to the mid-term examines and the break period, indicating that the rules can capture people's routine behavior. Rule names on the left indicate some example rules that are significantly different between the two classes of participants (see Table 3).

break period, some rules, which otherwise match many students, match very few students (represented by dark areas in the middle-top of the heatmap). This is positive evidence that *contextually filtered features capture routine behavior*, unlike their unimodal counterparts.

We further investigate the rules that capture different behavior patterns between students with depressive symptoms and students without depressive symptoms. We used a paired t-test on every rule to identify the rules that were significantly different between the two groups. Table 3 summarizes a subset of the top 20 rules in weekday/weekend rules that show the strongest significant difference (see the full list of top rules in Appendix).



(a) A Rule Common in Groups. More in Non-dep Group

(b) A Rule that is Unique to Non-dep Group

(c) A Rule Common in Groups. More in Dep Group

(d) A Rule that is Unique to in Dep Group

(e) A Similar Rule with Same Context Unique To Non-dep Group

(f) A Similar Rule with Same Context Unique To Dep Group

Fig. 4. Heatmaps of example rules that captures behavior differences between depression group and non-depression group. The percentage indicates the average proportion of students fulfilling the rule throughout the study period. Color indicates proportion of students on each day (X axis) Table 3 summarizes the details of the rules.

Weekday night rule No.5 (the first row in Table 3) indicates that students are be likely to have good sleep quality when they are on-campus and have low co-location (*i.e.*, the number of Bluetooth encounters) during weekday nights. This rule is present in both groups, but appears significantly more in the non-depression group ($t_{75} = 3.99, p < 0.001$, see Figure 4a). Weekday night rule No.9 (the second row in Table 3) indicates students' sleep bouts (periods of continuous sleep) are likely to be longer when they are on-campus and sleep efficiency is high. This rule is unique to the non-depression group ($t_{75} = 2.88, p < 0.01$, see Figure 4b).

Weekend morning rule No.3 (the third row in Table 3) indicates that when students have poor sleep (the sleep is intermittent), they are more likely to have low mobility (few location transitions) during weekend mornings (6am - 12pm). This rule is found in both groups, with significantly more students in the depression group ($t_{29} = -2.54, p < 0.05$, see Figure 4c) experiencing this. The *CondDisc* also shows that students with depressive symptoms are more like to match this rule's context (0.36 vs. 0.27), indicating worse sleep quality.

Another example rule reflects the relationship of mobile phone usage, sleep duration and depression. Weekend night rule No. 19 (the fourth row in Table 3) is only ranked high enough to be selected in the depression group. This suggests the potential effect of phone usage on sleep quality for depressive students. We discuss these findings more in Section 6.

We also find interesting unique rules in each group that reflected the differences between the two groups. Weekend morning rules No. 10 and No.11 (the last two rows in Table 3) share the same context set X, but have different Ys. Rule No. 10 is unique to the non-depression group and rule No. 11 is unique to the depression group. They indicated that students without depressive symptoms are more likely to have a high sleep efficiency when their location movement is low during weekend morning periods, but students with depressive symptoms are more likely to only have a medium sleep efficiency for the same context (see Figure 4e and 4f).

Table 3. Examples of rules that capture behavior difference between students with and without depressive symptoms. We tested a rule's ability to differentiate between classes using a paired t-test; significance level is indicated in the *Rule* column. We selected rules for this table that show the strongest significant difference. All top 20 weekday/weekend rules can be found in the appendix. *Type* is the method by which the rule was found. *Prop in Non-dep* and *Prop in Dep* are the proportion of students in a class that fulfill the rule, averaged over days in the study. Note that M varies between different epochs (people can have different behavior pattern during the day) as well as their types (*i.e.*, common or unique). *E.g.*, M of a weekday night rule can be much bigger than that of a weekday morning rule.

Rule	*X*	*Y*	Type	Prop in Non-dep	Prop in Dep	*Ctx Spec*	*Conf Diff*	*Cond Disc*	M
Wkdy Night No.5***	- [CampusMap] Percentage of time off-campus (low) - [Bluetooth] Number of unique device of others (low)	[Sleep] Sleep efficiency (high)	Common	11.4%	8.5%	2	0.137	0.094	0.031
Wkdy Night No.9**	- [CampusMap] Percentage of time off-campus (low) - [Sleep] Sleep efficiency (high)	[Sleep] Maximum length of asleep bouts (high)	Unique (Non-dep)	10.1%	7.9%	2	0.535	0.335	0.453
Wkend Morning No.3*	- [Sleep] Number of bouts being asleep (high) - [Sleep] Number of bouts being restless (high)	[Location] Number of location transition (low)	Common	9.2%	12.3%	2	0.054	0.081	0.007
Wkend Night No.19*	- [Screen] Mean length of screen being unlock (high)	[Sleep] Mean length of being asleep (low)	Unique (Dep)	7.8%	10.3%	1	0.387	0.374	0.147
Wkend Morning No.10***	- [Location] Number of location transition (low) - [CampusMap] Number of building transition on-campus (low)	[Sleep] Sleep efficiency (high)	Unique (Non-dep)	11.2%	5.5%	2	0.456	0.469	0.422
Wkend Morning No.11***	- [Location] Number of location transition (low) - [CampusMap] Number of building transition on-campus (low)	[Sleep] Sleep efficiency (medium)	Unique (Dep)	9.8%	16.6%	2	0.435	0.483	0.398

* indicates $p < 0.05$, ** indicates $p < 0.01$, *** indicates $p < 0.001$

5.2 Contextually Filtered Features Lead to Higher Performing Machine Learning Models

In this section, we show that the models trained on the contextually filtered features extracted via our method can achieve a better performance than other models. We also perform an ablation study on the three components in metric M (*CtxSpec*, *ConfDiff*, *CondDisc*), which demonstrates the relative importance of the three characteristics as *ConfDiff* > *CtxSpec* > *CondDisc*. We validate the generalizability of our method by applying it to separate dataset (the Phase II data described in Section 4.1).

5.2.1 Contextually Filtered Features Can Better Identify Students with Depressive Symptoms. Recall that after we obtained the best rules from our RuleGenerateSet (50 students: 20 in the depression group and 30 in the non-depression group), we extracted contextually filtered features using these rules on the TrainTestSet (88 students, 37 in the depression group and 51 in the non-depression group).

Table 4. Comparison of baseline machine learning classifiers and contextually filtered features. The models above the dashed line are baselines. Models based on unimodal features, contextually filtered features, and hybrid features, are trained using AdaBoost [35] with decision-tree-based component classifiers, with leave-one-out cross-validation. The number of estimator and the maximum depth of the decision-tree are hyper-parameters that can be tuned. We use grid search to select the best parameters for each model. Our best model with hybrid features has the number of estimator as 10 and the maximum depth as 3. A t-test on the test results between the hybrid features and unimodal raw features show that our method significantly outperforms the standard method ($p < 0.01$).

Classification	Features	Accuracy	Precision	Recall	F1 Score
	Majority	0.579	0.579	1.000	0.734
	Best Single Feature	0.704	0.725	0.755	0.740
CPAR [95]	Class Association Rules	0.608	0.629	0.850	0.723
AdaBoost [35]	Unimodal Features	0.716	0.725	0.771	0.747
AdaBoost [35]	Contextually Filtered Features	0.807	0.765	**0.886**	0.821
AdaBoost [35]	Hybrid Features	**0.818**	**0.843**	0.843	**0.843**
	Performance Increase of Hybrid over Unimodal	*10.2%*	*11.8%*	*7.2%*	*9.6%*
			Average Increase:		***9.7%***

We tested two feature sets: 1) **Contextually Filtered Features**: only the features extracted based on rules (vector length 274); 2) **Hybrid Features**: both the contextually filtered features and the unimodal features (vector length 455 (274+181), see Section 4.3.2). We employed AdaBoost [35] with decision-tree-based component classifiers during the training. To avoid over-fitting, we used leave-one-out cross-validation, since previous work has consistently found that this method is approximately unbiased and has small variance [79, 96].

We compared our models with four baselines: 1) Majority: the classifier simply predicts the major label in the dataset (*i.e., no depressive symptoms*; 2) Best Single Feature: prediction is made based on the value of the single feature that best distinguishes the classes; 3) Class Association Rules: labels are embedded into the input during association rule mining and the generated rules are used for classification [56, 95]; 4) Unimodal Features: the model is trained on the unimodal features before rule mining (vector length 181, a common practice in previous work [84]).

We summarize the results in Table 4 with four metrics: accuracy, precision, recall and F1 score. The model trained on the hybrid features has the best performance, followed by the model trained on contextually filtered features. Our best model has accuracy 0.818 and F1 score 0.843. It outperforms the baseline model using the unimodal features, by an average of 9.7% absolute increase, indicating the effectiveness of our method. Since the area of using mobile sensing for depression detection is fairly new, we lack a benchmark for comparison. However, these baselines provide strong evidence that our model is either better than previous work [19, 30, 84], or comparable to the state-of-the-art [70, 87].

5.2.2 Relative Importance of The Three Characteristics For Classification. M (Equation 1) is composed of three characteristics: Contextual Specificity, Confidence Difference and Condition Discrepancy. It is interesting to examine which component is important for rule selection, so that we can have a better understanding of metric M.

The weight values, calculated using grid search in Section 3, reflect the relative importance of the three characteristics. The greater the weight value is, the more important role the corresponding characteristic plays in the metric M. Our weights show the importance of *ConfDiff* ($w_2 = 1.5$), followed by *CtxSpec* ($w_1 = 1.0$), followed by *CondDisc* ($w_3 = 0.5$).

Table 5. Results of the ablation study. One of the three weight values is set to zero in each trial, which can lead to different rule sets, and new models are trained based on these rules. The other weights (w_1, w_2, w_3) are set to $(1.0, 1.5, 0.5)$, as described in Section 3. The results are presented in an ascending order according to F1 score.

Classification	Ablated Metric	Accuracy	Precision	Recall	F1 Score
Depression Detection with Contextually Filtered Features	*ConfDiff* - $w_2 = 0$	0.761	0.804	0.788	0.796
	CtxSpec - $w_1 = 0$	0.761	0.863	0.759	0.807
	CondDisc - $w_3 = 0$	0.784	0.808	0.824	0.816
	No Metric Ablated	0.807	0.765	0.886	0.821

We further examined the effect of each characteristic, using an ablation study on the three weights. We set one of the weights to zero in each trial and redo the rule selection, feature extraction and modeling training. Table 5 summarizes the results. Removing *CtxSpec* $(w_1 = 0)$ and *ConfDiff* $(w_2 = 0)$ lead to similar results, with both models having a drop in accuracy of 4.5 percentage-points. The model without *CtxSpec* has a slightly higher F1 score than without *ConfDiff*. Removing *CondDisc* $(w_3 = 0)$ has the least impact on the results, with two percentage-points drop in accuracy. These results are consistent with the relative order of weight values. Confidence Difference is the most essential part in the metric M, and the Condition Discrepancy is the least important part.

5.3 Verification on A Second Dataset

Conducting such a large-scale data collection study (as described in Section 4) can be very expensive in terms of time and money. Despite this, knowing how well our method can perform on another dataset can tell us about its generalizability.

We collected a separate Phase II dataset one year later from the same university (as described in Section 4.1). Of the 211 participants with good data in Phase II, 65 also had participated in the Phase I study. The same data collection apps and wearable devices were used in the two phases. This provides a unique opportunity to verify our method in a consistent way.

There are three aspects to the robustness of our method: 1) model-level, 2) rule-level, and 3) pipeline-level. Most existing work tests robustness using cross-validation in which training and testing are an average of iterative trials run on a single data divided into train and test data. Our work does this as well. However, unlike all of the past work we have been able to find, we have the opportunity to also test our work on multiple data sets. This lets us test several forms of robustness:

(1) To study model-level robustness, we run the whole pipeline on Phase I dataset, train the model on the Phase I dataset, and test the model on the Phase II dataset.
(2) To study rule-level robustness, we split the pipeline into two datasets, mine the rules from the Phase I dataset, use these rules to extract contextually filtered feature on Phase II dataset, and train/test the model on the Phase II dataset.
(3) To study pipeline-level robustness, we replicate the whole pipeline on Phase II.

In this work, we test all 3 forms of robustness. Table 6 summarizes the results of 2) rule-level and 3) pipeline-level robustness. We also tested (1) model-level robustness. However, our model is not reliable on the second dataset (accuracy of 54.2%, no better than a majority-based baseline predictor).

5.3.1 Verification of The Generalizability of Rules.

We select rules using Phase I data. We then use the rules on the the Phase II dataset to extract contextually filtered features and train the models. The top half of Table 6 summarizes the results. Despite the similarities in the data collection and in the student population, we expect a drop in the performance from Phase I (see Table 4), due to not having the exact same participants and to Phase II

Table 6. Verification results. The same model and cross-validation technique are used as in Table 4. A paired t-test comparing the hybrid features and unimodal features shows strong significance, for both the rule and pipeline verification ($p < 0.001$).

Classification	Features	Accuracy	Precision	Recall	F1 Score
Verification On Phase II Dataset with The Rules From Phase I	Majority	0.643	0.643	1.000	0.783
	Best Single Feature	0.646	0.655	0.949	0.775
	Class Association Rules	0.601	0.642	0.854	0.733
	Unimodal Features	0.656	0.701	0.809	0.751
	Contextually Filtered Features	0.689	0.757	0.779	0.768
	Hybrid Features	**0.731**	**0.762**	**0.846**	**0.801**
	Performance Increase of Hybrid over Unimodal	*7.5%*	*6.1%*	*3.7%*	*5.0%*
			Average Increase:		***5.6%***
Verification with The Pipeline on Phase II Dataset	Majority	0.656	0.656	1.000	0.793
	Best Single Feature	0.702	0.745	0.824	0.782
	Class Association Rules	0.626	0.804	0.691	0.743
	Unimodal Features	0.740	0.760	0.884	0.817
	Contextually Filtered Features	0.809	**0.877**	0.826	0.850
	Hybrid Features	**0.840**	0.857	**0.907**	**0.881**
	Performance Increase of Hybrid over Unimodal	*10.0%*	*9.7%*	*2.3%*	*6.4%*
			Average Increase:		***7.1%***

occurring one year after Phase I. Indeed, the best model has accuracy 0.731 and F1 score 0.801 (compared to 0.818 and 0.843, respectively one Phase I only).

In addition, our model still outperforms all the baselines on the Phase II dataset. It also outperforms the model built with the unimodal features by an average of 5.6% absolute increase on the metrics of accuracy, precision, recall and F1 score. Baselines were all prepared using only Phase II data, making these results all the more impressive. These results verify the generalizability and overall stability of the outcome rules from our method.

5.3.2 Verification of The Generalizability of Pipeline. As an additional verification, we reapply the whole pipeline as described in Section 4.3 on Phase II. We omit the grid search and set the weights as $w_1 = 1.0, w_2 = 1.5, w_3 = 0.5$, since M, as a general formula capturing interesting rules, should be the same on either dataset. The bottom half of Table 6 summarizes the results. The best model (hybrid of unimodal and contextually filtered features) has an accuracy of 0.840 and F1 score of 0.881.

This pipeline again outperforms the baseline models, which are also trained entirely on Phase II. This model also outperforms the unimodal features model by an average of 7.1% absolute increase on the metrics. These results validate the generalizability of our overall algorithm.

6 DISCUSSION

In this section, we discuss insights obtained from our analysis and implications for intervention design for depression. We also discuss potential directions for generalizing and improving our approach.

6.1 Relation to the Depression Literature

Our findings in Section 5 are consistent with the current literature on depression, adding support for the validity of our methods. For example, the features in Y in weekday night rule No.5 and No.9 (see top 2 rules in Table 3) suggest that those students with depressive symptoms are less likely to have good sleep quality (high sleep efficiency and long asleep bouts). The contrast between weekend morning rule No.10 and No.11 (see bottom 2 rules in Table 3) also reveal this for students who are in the same context. These results can be supported by relevant findings in psychology and clinical psychiatry that sleep disturbance is a common symptom of depression [7, 78, 80]. Weekend morning rule No.3 implies a relationship between depression and both mobility and sleep, that not only echoes previous literature regarding the effect of depression on sleep [7, 87], but also is supported by the findings of other studies similar to ours which show that depression and diminished locomotion co-occur [19, 70]. Weekend night rule No. 19 suggests the potential effect of phone usage on sleep quality for students with symptoms of depression. Although this rule does not show a direct relation between phone usage and depression, it does reflect the rich literature that depression may lead to more phone usage [24, 36, 70, 84].

6.1.1 Location and Sleep Information for Depression Detection. The top rules for feature extraction and model training cover all type of sensors (except phone calls) in Figure 1, showing the multimodality of our method. The absence of calls in these rules might be explained by the fact that an increasing number of students use social media platforms or text messages, instead of phone calls, for communication, resulting in less informative data in the call logs. Among the 105 top rules, we observed a large number of rules involving location and sleep: 89 rules had at least one feature (in either X or Y) relevant to *Location*, 75 rules had at least one feature relevant to *CampusMap* which is actually based on *Location*, and 50 rules had at least one feature relevant to *Sleep*. Examples in Table 3 also reveal the dominance of location and sleep information in the rules. This resonates with findings in other work about mobile sensing for depression detection [19, 70, 84, 87].

6.2 Robustness and Generalizability of Our Method

Our results demonstrate strong robustness at the rule and pipeline level. Our approach is significantly better than baseline models on the Phase II data in both cases. To our knowledge, no prior work has explored this issue and our dataset is unique in allowing for multi-year robustness verification.

We found that model robustness is not as reliable (accuracy of 54.2%, no better than a majority-based baseline predictor). An important area of future work will be the development of modeling approaches that are robust over multiple years and in new student populations.

6.3 Beyond Depression Detection

Our method in Section 3 is agnostic to the specific classes on which it is trained. In this paper, we focus on depression detection among college students, and split the dataset based on student scores on the BDI-II, which indicate the presence of depressive symptoms. It would be interesting to explore other prediction tasks. For instance, instead of focusing on detecting which students in our population will have symptoms of depression at the end of the semester, we could focus on detecting which students are successfully coping with their depressive symptoms by maintaining or improving their BDI-II score over the course of the semester, and which students are experiencing more severe depressive symptoms at the end of the semester. This could be explored by splitting based on the direction of change in the BDI-II score from pre-semester to post-semester, for those students with medium to high BDI-II scores at the start of the semester.

Further, our method can be applied outside the domain of depression, and to other time-series datasets about human behavior, to detect behaviors and states of interest in the studied populations. Compared to previous work such as [87], our method does not depend on domain knowledge and hand-crafted features.

One open problem for our approach is how to generalize it to multi-class rather than two class problems. While this should be a straightforward extension of our rule selection methods, it remains as future work.

6.4 Leveraging Association Rule Mining and Other Algorithms

There are a number of metrics for selecting rules mined using ARM that are not covered in this paper (lift [59], match [89], *etc.*). We heuristically designed our metric M using criteria that capture differences between two groups. It has some space for improvement. For instance, the current M will rank a rule with high context probability ($P(X)$) but different signs ($DirDiff = 0$) at bottom, which may miss some interesting information. More complex metrics can be explored based on the outcomes of traditional ARM. Recent work such as temporal association rule mining [44, 85] and graph association rule mining [14, 60], have the potential to take temporal information into account. In addition, sequential pattern mining [15] and sequential rule mining [34] can also be employed to investigate temporal sequences or behavior sequences. Moreover, some deep-learning techniques such as long short-term memory (LSTM) [41] may capture more nonlinear and complex relationships among features in the neural network. Although current deep-learning models are relatively less interpretable, more and more works try to understand the principle of neural network [48]. These approaches all have the potential to be combined with our method to identify contextualized behavior differences between groups, providing richer information to understand human behavior.

7 LIMITATIONS

In this section, we describe a few of the limitations of our work. First, we only had the post-semester BDI-II score to use as ground truth, resulting in a single label per student over the whole semester. As such, compared to other work such as [87], we were not able to investigate more fine-grained dynamics of students' behavior. In the future, we plan to collect depression scores more frequently to support a more fine-grained analysis. Second, in the Phase II dataset, we did not explicitly remove participants who also participated in the Phase I dataset. This could affect the results of Section 5.3.2, where we mined rules from Phase I and tested it on Phase II, since there were overlapping students in both datasets. While the re-application of the entire pipeline on Phase II did demonstrate the generalizability of our approach, separating out the repeat participants could help in better understanding the generalizability of the rules extracted from Phase I. Third, our method relies on the unimodal features extracted from the dataset. The rule mining is applied on the unimodal features. Thus the capability of our method is limited by these features. If the unimodal features do not capture any aspect of users' behavior, neither can our method do. There may exist more meaningful features to be extracted at the unimodal feature extraction stage (see Section 4.2), which may enable our method to better capture behavior routines as well as behavior pattern differences. Finally, further methods for dealing with missing data could be explored. We removed user-days points that were missing more than half of the features from the study to avoid the bias of low-quality data. But this might neglect the case where a day of missing data could be related to students' depression status (*e.g.*, not charging the phone because of the diminished desire for social interaction [7]). However, the percentage of students with depressive symptoms on the BDI-II who were removed from the dataset due to missing data (8 out of 24) is similar to the percentage who were not removed, which lends us confidence that our data is representative even after those students were removed.

8 CONCLUSION

In this paper, we present a new method based on association rule mining for generating **contextually filtered features** in an automated way, which can perform better than standard feature selection approaches for depression detection. We apply our novel method on a passive mobile and wearable dataset with 138 college students, whose depressive symptoms at the end of the semester were measured by their post-semester BDI-II score. We show that the best rules selected by our method are highly interpretable and can capture students' routine behaviors, and behavior pattern differences between students with and without depressive symptoms. Based on the resulting **contextually filtered features**, we train classifiers to predict whether a student will have depressive symptoms at the end of the semester (*i.e.*, the post-semester BDI-II score greater than 13) based on their behavior during the semester. We demonstrate that our best model outperforms a standard model by an average of 9.7% across a variety of metrics. We further verify the generalizability of our method by applying both the rules from the original dataset, and the overall method, to a second similar dataset. Our best model outperforms the standard approach by an average of 5.6% and 7.1%, respectively.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grant Number IIS1816687 and IIS7974751, the National Institute on Disability, Independent Living and Rehabilitation Research under Grant Number 90DPGE0003-01, Carnegie Mellon University Office of the Provost, University of Washington College of Engineering, and an Adobe Data Science Research Grant.

REFERENCES

[1] Gregory D. Abowd, Anind K. Dey, Peter J. Brown, Nigel Davies, Mark Smith, and Pete Steggles. 1999. Towards a better understanding of context and context-awareness. In *International Symposium on Handheld and Ubiquitous Computing*. Springer, 304–307.

[2] Substance Abuse, Mental Health Services Administration, et al. 2016. 2015 National Survey on Drug Use and Health. (2016).

[3] ACHA-NCHA II 2018. Undergraduate Student Reference Group - Data Report. https://www.acha.org/documents/ncha/NCHA-II_Spring_2018_Undergraduate_Reference_Group_Data_Report.pdf. [Online; accessed 19-July-2008].

[4] Rakesh Agrawal and Ramakrishnan Srikant. 1994. Fast algorithms for mining association rules. In *Proceedings of 20th International Conference on Very Large Data Bases, VLDB*, Vol. 1215. 487–499.

[5] Rakesh Agrawal and Ramakrishnan Srikant. 1995. Mining sequential patterns. In *Proceedings of the 11th International Conference on Data Engineering*. IEEE, 3–14.

[6] Maria-Luiza Antonie, Osmar R. Zaiane, and Alexandru Coman. 2001. Application of data mining techniques for medical image classification. In *Proceedings of the Second International Conference on Multimedia Data Mining*. Springer-Verlag, 94–101.

[7] American Psychiatric Association et al. 2013. *Diagnostic and statistical manual of mental disorders (DSM-5®).* American Psychiatric Pub.

[8] Min S. Hane Aung, Faisal Alquaddoomi, Cheng-Kang Hsieh, Mashfiqui Rabbi, Longqi Yang, John P. Pollak, Deborah Estrin, and Tanzeem Choudhury. 2016. Leveraging multi-modal sensing for mobile health: a case review in chronic pain. *IEEE Journal of Selected Topics in Signal Processing* 10, 5 (2016), 962–974.

[9] Nikola Banovic, Tofi Buzali, Fanny Chevalier, Jennifer Mankoff, and Anind K. Dey. 2016. Modeling and understanding human routine behavior. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 248–260.

[10] Nikola Banovic, Anqi Wang, Yanfeng Jin, Christie Chang, Julian Ramos, Anind K. Dey, and Jennifer Mankoff. 2017. Leveraging human routine models to detect and generate human behaviors. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM, 6683–6694.

[11] Aaron T. Beck. 1979. *Cognitive therapy of depression.* Guilford press.

[12] Aaron T. Beck, Robert A. Steer, and Gregory K. Brown. 1996. Beck depression inventory-II. *San Antonio* 78, 2 (1996), 490–498.

[13] Dror Ben-Zeev, Emily A. Scherer, Rui Wang, Haiyi Xie, and Andrew T. Campbell. 2015. Next-generation psychiatric assessment: Using smartphone sensors to monitor behavior and mental health. *Psychiatric Rehabilitation Journal* 38, 3 (2015), 218.

[14] Michele Berlingerio, Francesco Bonchi, Björn Bringmann, and Aristides Gionis. 2009. Mining graph evolution rules. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 115–130.

[15] Oliver Brdiczka, Norman Makoto Su, and Bo Begole. 2009. Using temporal patterns (t-patterns) to derive stress factors of routine tasks. In *CHI'09 Extended Abstracts on Human Factors in Computing Systems*. ACM, 4081–4086.

[16] Oliver Brdiczka, Norman Makoto Su, and James Bo Begole. 2010. Temporal task footprinting: identifying routine tasks by their temporal patterns. In *Proceedings of the 15th International Conference on Intelligent User Interfaces*. ACM, 281–284.

[17] George W. Brown, Bernice Andrews, Tirril Harris, Zsuzsanna Adler, and L. Bridge. 1986. Social support, self-esteem and depression. *Psychological medicine* 16, 4 (1986), 813–831.

[18] Michelle Nicole Burns, Mark Begale, Jennifer Duffecy, Darren Gergle, Chris J. Karr, Emily Giangrande, and David C. Mohr. 2011. Harnessing context sensing to develop a mobile intervention for depression. *Journal of Medical Internet Research* 13, 3 (2011).

[19] Luca Canzian and Mirco Musolesi. 2015. Trajectories of depression: unobtrusive monitoring of depressive states by means of smartphone mobility traces analysis. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous computing*. ACM, 1293–1304.

[20] Huanhuan Cao, Tengfei Bao, Qiang Yang, Enhong Chen, and Jilei Tian. 2010. An effective approach for mining mobile user habits. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*. ACM, 1677–1680.

[21] Philip I. Chow, Karl Fua, Yu Huang, Wesley Bonelli, Haoyi Xiong, Laura E. Barnes, and Bethany A. Teachman. 2017. Using mobile sensing to test clinical models of depression, social anxiety, state affect, and social isolation among college students. *Journal of Medical Internet Research* 19, 3 (2017).

[22] Ewa K. Czyz, Adam G. Horwitz, Daniel Eisenberg, Anne Kramer, and Cheryl A. King. 2013. Self-reported barriers to professional help seeking among college students at elevated risk for suicide. *Journal of American College Health* 61, 7 (2013), 398–406.

[23] Antonio Reis de Sá Junior, Arthur Guerra de Andrade, Laura Helena Andrade, Clarice Gorenstein, and Yuan-Pang Wang. 2018. Response pattern of depressive symptoms among college students: What lies behind items of the Beck Depression Inventory-II? *Journal of Affective Disorders* 234 (2018), 124–130.

[24] Kadir Demirci, Mehmet Akgönül, and Abdullah Akpinar. 2015. Relationship of smartphone use severity with sleep quality, depression, and anxiety in university students. *Journal of Behavioral Addictions* 4, 2 (2015), 85–92.

[25] Anind K. Dey. 2001. Understanding and using context. *Personal and Ubiquitous Computing* 5, 1 (2001), 4–7.

[26] Afsaneh Doryab. 2018. Identifying symptoms using technology. In *Technology and Adolescent Mental Health*. Springer, 135–153.

[27] Afsaneh Doryab, Jun-Ki Min, Jason Wiese, John Zimmerman, and Jason I. Hong. 2014. Detection of behavior change in people with depression. In *AAAI Workshop: Modern Artificial Intelligence for Health Analytics*.

[28] 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.

[29] Daniel Eisenberg, Ezra Golberstein, and Sarah E Gollust. 2007. Help-seeking and access to mental health care in a university student population. *Medical Care* (2007), 594–601.

[30] Asma Ahmad Farhan, Chaoqun Yue, Reynaldo Morillo, Shweta Ware, Jin Lu, Jinbo Bi, Jayesh Kamath, Alexander Russell, Athanasios Bamis, and Bing Wang. 2016. Behavior vs. introspection: refining prediction of clinical depression via smartphone sensing data. In *Wireless Health*. 30–37.

[31] Katayoun Farrahi and Daniel Gatica-Perez. 2012. Extracting mobile behavioral patterns with the distant n-gram topic model. In *Proceedings of the 16th International Symposium on Wearable Computers (ISWC)*. IEEE, 1–8.

[32] Denzil Ferreira, Vassilis Kostakos, and Anind K. Dey. 2015. AWARE: mobile context instrumentation framework. *Frontiers in ICT* 2 (2015), 6.

[33] Philippe Fournier-Viger, Antonio Gomariz, Ted Gueniche, Azadeh Soltani, Cheng-Wei Wu, and Vincent S. Tseng. 2014. SPMF: a Java open-source pattern mining library. *The Journal of Machine Learning Research* 15, 1 (2014), 3389–3393.

[34] Philippe Fournier-Viger, Ted Gueniche, Souleymane Zida, and Vincent S Tseng. 2014. ERMiner: sequential rule mining using equivalence classes. In *International Symposium on Intelligent Data Analysis*. Springer, 108–119.

[35] Yoav Freund and Robert E. Schapire. 1997. A decision-theoretic generalization of on-line learning and an application to boosting. *J. Comput. System Sci.* 55, 1 (1997), 119–139.

[36] Mads Frost, Afsaneh Doryab, Maria Faurholt-Jepsen, Lars Vedel Kessing, and Jakob E. Bardram. 2013. Supporting disease insight through data analysis: refinements of the monarca self-assessment system. In *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 133–142.

[37] Toshi A Furukawa. 2010. Assessment of mood: guides for clinicians. *Journal of Psychosomatic Research* 68, 6 (2010), 581–589.

[38] Liqiang Geng and Howard J. Hamilton. 2006. Interestingness measures for data mining: a survey. *Comput. Surveys* 38, 3, Article 9 (Sept. 2006). https://doi.org/10.1145/1132960.1132963

[39] D. Goldberg, K. Bridges, P. Duncan-Jones, and D. Grayson. 1988. Detecting anxiety and depression in general medical settings. *British Medical Journal* 297, 6653 (1988), 897–899.

[40] Darcy Gruttadaro and Dana Crudo. 2012. College students speak: A survey report on mental health. *National Alliance on Mental Illness* (2012).

[41] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation* 9, 8 (1997), 1735–1780.

[42] Alketa Hysenbegasi, Steven L Hass, and Clayton R Rowland. 2005. The impact of depression on the academic productivity of university students. *Journal of Mental Health Policy and Economics* 8, 3 (2005), 145.

[43] Varun Jain, James L Crowley, Anind K. Dey, and Augustin Lux. 2014. Depression estimation using audiovisual features and fisher vector encoding. In *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*. ACM, 87–91.

[44] Thomas Janssoone, Chloé Clavel, Kévin Bailly, and Gaël Richard. 2016. Using temporal association rules for the synthesis of embodied conversational agents with a specific stance. In *International Conference on Intelligent Virtual Agents*. Springer, 175–189.

[45] Szymon Jaroszewicz and Dan A Simovici. 2001. A general measure of rule interestingness. In *European Conference on Principles of Data Mining and Knowledge Discovery*. Springer, 253–265.

[46] Richard Kadison and Theresa Foy DiGeronimo. 2004. College of the overwhelmed: The campus mental health crisis and what to do about it. *San Francisco* (2004).

[47] Yasutaka Kamei, Akito Monden, Shuji Morisaki, and Ken-ichi Matsumoto. 2008. A hybrid faulty module prediction using association rule mining and logistic regression analysis. In *Proceedings of the Second ACM-IEEE International Symposium on Empirical Software Engineering and Measurement*. ACM, 279–281.

[48] Andrej Karpathy, Justin Johnson, and Li Fei-Fei. 2015. Visualizing and understanding recurrent networks. *arXiv preprint arXiv:1506.02078* (2015).

[49] Raghavendra Katikalapudi, Sriram Chellappan, Frances Montgomery, Donald Wunsch, and Karl Lutzen. 2012. Associating internet usage with depressive behavior among college students. *IEEE Technology and Society Magazine* 31, 4 (2012), 73–80.

[50] Ronald C. Kessler and Evelyn J Bromet. 2013. The epidemiology of depression across cultures. *Annual Review of Public Health* 34 (2013), 119–138.

[51] Keivan Kianmehr and Reda Alhajj. 2006. Effective classification by integrating support vector machine and association rule mining. In *International Conference on Intelligent Data Engineering and Automated Learning*. Springer, 920–927.

[52] Jeremy Kisch, E Victor Leino, and Morton M Silverman. 2005. Aspects of suicidal behavior, depression, and treatment in college students: Results from the Spring 2000 National College Health Assessment Survey. *Suicide and Life-Threatening Behavior* 35, 1 (2005), 3–13.

[53] Sotiris Kotsiantis and Dimitris Kanellopoulos. 2006. Association rules mining: A recent overview. *GESTS International Transactions on Computer Science and Engineering* 32, 1 (2006), 71–82.

[54] Alexander Kraskov, Harald Stögbauer, and Peter Grassberger. 2004. Estimating mutual information. *Physical Review E* 69, 6 (2004), 066138.

[55] Nicholas D. Lane, Emiliano Miluzzo, Hong Lu, Daniel Peebles, Tanzeem Choudhury, and Andrew T. Campbell. 2010. A survey of mobile phone sensing. *IEEE Communications Magazine* 48, 9 (2010).

[56] Bing Liu, Wynne Hsu, and Yiming Ma. 1998. Integrating classification and association rule mining. In *Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining*.

[57] Guimei Liu, Haojun Zhang, and Limsoon Wong. 2014. A flexible approach to finding representative pattern sets. *IEEE Transactions on Knowledge and Data Engineering* 26, 7 (2014), 1562–1574.

[58] Magnus S. Magnusson. 2000. Discovering hidden time patterns in behavior: T-patterns and their detection. *Behavior Research Methods, Instruments, & Computers* 32, 1 (2000), 93–110.

[59] Paul David McNicholas, Thomas Brendan Murphy, and M. O'Regan. 2008. Standardising the lift of an association rule. *Computational Statistics & Data Analysis* 52, 10 (2008), 4712–4721.

[60] Mohammad Hossein Namaki, Yinghui Wu, Qi Song, Peng Lin, and Tingjian Ge. 2017. Discovering graph temporal association rules. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. ACM, 1697–1706.

[61] Suman Nath. 2012. ACE: exploiting correlation for energy-efficient and continuous context sensing. In *Proceedings of the 10th International Conference on Mobile Systems, Applications, and Services*. ACM, 29–42.

[62] Sarfraz Nawaz and Cecilia Mascolo. 2014. Mining users' significant driving routes with low-power sensors. In *Proceedings of the 12th ACM Conference on Embedded Network Sensor Systems*. ACM, 236–250.

[63] NIMH Website 2018. Depression - National Institute of Mental Health. https://www.nimh.nih.gov/health/topics/depression/index.shtml

[64] Matthew K. Nock and Ronald C. Kessler. 2006. Prevalence of and risk factors for suicide attempts versus suicide gestures: analysis of the National Comorbidity Survey. *Journal of Abnormal Psychology* 115, 3 (2006), 616.

[65] Jong Soo Park, Ming-Syan Chen, and Philip S. Yu. 1997. Using a hash-based method with transaction trimming for mining association rules. *IEEE Transactions on Knowledge and Data Engineering* 9, 5 (1997), 813–825.

[66] Emma Pierson, Tim Althoff, and Jure Leskovec. 2018. Modeling individual cyclic variation in human behavior. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 107–116.

[67] J. Ross Quinlan. 1986. Induction of decision trees. *Machine Learning* 1, 1 (1986), 81–106.

[68] Periyasamy Rajendran and Muthusamy Madheswaran. 2010. Hybrid medical image classification using association rule mining with decision tree algorithm. *arXiv preprint arXiv:1001.3503* (2010).

[69] Sohrab Saeb, Emily G. Lattie, Stephen M. Schueller, Konrad P. Kording, and David C. Mohr. 2016. The relationship between mobile phone location sensor data and depressive symptom severity. *PeerJ* 4 (2016), e2537.

[70] Sohrab Saeb, Mi Zhang, Christopher J. Karr, Stephen M. Schueller, Marya E. Corden, Konrad P. Kording, and David C. Mohr. 2015. Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: an exploratory study. *Journal of Medical Internet Research* 17, 7 (2015).

[71] Iqbal H. Sarker and Flora D. Salim. 2018. Mining user behavioral rules from smartphone data through association analysis. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 450–461.

[72] Vijay Srinivasan, Christian Koehler, and Hongxia Jin. 2018. RuleSelector: Selecting conditional action rules from user behavior patterns. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies* 2, 1 (2018), 35.

[73] Vijay Srinivasan, Saeed Moghaddam, Abhishek Mukherji, Kiran K Rachuri, Chenren Xu, and Emmanuel Munguia Tapia. 2014. Mobileminer: Mining your frequent patterns on your phone. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 389–400.

[74] Robert A. Steer, Gregory. K Brown, Aaron T. Beck, and William C. Sanderson. 2001. Mean Beck Depression Inventory-II scores by severity of major depressive episode. *Psychological Reports* 88, 3_suppl (2001), 1075–1076.

[75] Eric A. Storch, Jonathan W. Roberti, and Deborah A. Roth. 2004. Factor structure, concurrent validity, and internal consistency of the beck depression inventory-second edition in a sample of college students. *Depression and Anxiety* 19, 3 (2004), 187–189.

[76] Yoshihiko Suhara, Yinzhan Xu, and Alex 'Sandy' Pentland. 2017. Deepmood: Forecasting depressed mood based on self-reported histories via recurrent neural networks. In *Proceedings of the 26th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 715–724.

[77] Laszlo Szathmary. 2006. *Symbolic data mining methods with the Coron platform*. Ph.D. Dissertation. Université Henri Poincaré-Nancy I.

[78] Michael E. Thase. 1998. Depression, sleep, and antidepressants. *The Journal of Clinical Psychiatry* (1998).

[79] Lu Tian, Tianxi Cai, Els Goetghebeur, and LJ Wei. 2007. Model evaluation based on the sampling distribution of estimated absolute prediction error. *Biometrika* 94, 2 (2007), 297–311.

[80] Norifumi Tsuno, Alain Besset, and Karen Ritchie. 2005. Sleep and depression. *The Journal of Clinical Psychiatry* (2005).

[81] Michael Von Korff and Gregory Simon. 1996. The relationship between pain and depression. *The British Journal of Psychiatry* 168, S30 (1996), 101–108.

[82] Theo Vos and the GBD 2015 Disease and Injury Incidence and Prevalence Collaborators. 2016. Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990–2015: a systematic analysis for the Global Burden of Disease Study 2015. *The Lancet* 388, 10053 (2016), 1545–1602.

[83] Fabian Wahle, Lea Bollhalder, Tobias Kowatsch, and Elgar Fleisch. 2017. Toward the design of evidence-based mental health information systems for people with depression: A systematic literature review and meta-analysis. *Journal of Medical Internet Research* 19, 5 (2017), e191. https://doi.org/10.2196/jmir.7381

[84] Fabian Wahle, Tobias Kowatsch, Elgar Fleisch, Michael Rufer, and Steffi Weidt. 2016. Mobile sensing and support for people with depression: A pilot trial in the wild. *JMIR mHealth and uHealth* 4, 3 (2016), e111. https://doi.org/10.2196/mhealth.5960

[85] Ling Wang, Jianyao Meng, Peipei Xu, and Kaixiang Peng. 2018. Mining temporal association rules with frequent itemsets tree. *Applied Soft Computing* 62 (2018), 817–829.

[86] Rui Wang, Fanglin Chen, Zhenyu Chen, Tianxing Li, Gabriella Harari, Stefanie Tignor, Xia Zhou, Dror Ben-Zeev, and Andrew T. Campbell. 2014. StudentLife: assessing mental health, academic performance and behavioral trends of college students using smartphones. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 3–14.

[87] Rui Wang, Weichen Wang, Alex daSilva, Jeremy F. Huckins, William M. Kelley, Todd F. Heatherton, and Andrew T. Campbell. 2018. Tracking depression dynamics in college students using mobile phone and wearable sensing. *Proceedings of ACM Interactive, Mobile, Wearable and Ubiquitous Technology* 2, 1, Article 43 (March 2018), 26 pages. https://doi.org/10.1145/3191775

[88] X. Wang, A. McCallum, and X. Wei. 2007. Topical n-grams: Phrase and topic discovery, with an application to information retrieval. In *Seventh IEEE International Conference on Data Mining (ICDM 2007)*. 697–702. https://doi.org/10.1109/ICDM.2007.86

[89] Jin-Mao Wei, Wei-Guo Yi, and Ming-Yang Wang. 2005. Novel measurement for mining effective association rules. In *2005 International Conference on Machine Learning and Cybernetics*, Vol. 3. IEEE, 1660–1664.

[90] Wenmin Li, Jiawei Han, and Jian Pei. 2001. CMAR: accurate and efficient classification based on multiple class-association rules. In *Proceedings 2001 IEEE International Conference on Data Mining*. 369–376. https://doi.org/10.1109/ICDM.2001.989541

[91] Mark A Whisman, Charles M Judd, Natalie T Whiteford, and Heather L Gelhorn. 2013. Measurement invariance of the Beck Depression Inventory–Second Edition (BDI-II) across gender, race, and ethnicity in college students. *Assessment* 20, 4 (2013), 419–428.

[92] Mark A Whisman and Emily D Richardson. 2015. Normative data on the Beck Depression Inventory–second edition (BDI-II) in college students. *Journal of Clinical Psychology* 71, 9 (2015), 898–907.

[93] Dong Xin, Hong Cheng, Xifeng Yan, and Jiawei Han. 2006. Extracting redundancy-aware top-k patterns. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 444–453.

[94] Xifeng Yan, Hong Cheng, Jiawei Han, and Dong Xin. 2005. Summarizing itemset patterns: a profile-based approach. In *Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*. ACM, 314–323.

[95] Xiaoxin Yin and Jiawei Han. 2003. CPAR: Classification based on predictive association rules. In *Proceedings of the 2003 SIAM International Conference on Data Mining*. SIAM, 331–335.

[96] Yongli Zhang and Yuhong Yang. 2015. Cross-validation for selecting a model selection procedure. *Journal of Econometrics* 187, 1 (2015), 95–112.

[97] Brian D Ziebart. 2010. *Modeling purposeful adaptive behavior with the principle of maximum causal entropy*. Ph.D. Dissertation. Carnegie Mellon University.

APPENDIX

Table 7. Top 20 rules from the four epochs of weekdays that capture behavior difference between depression group and non-depression group. Type is the method by which the rule was found. *CtxSpec* indicates *contextual specificity* ($|X|$) of a rule. *ConfDiff* indicates *confidence difference* ($|\Delta conf|$) of a rule between two groups. *CondDisc* indicates *condition discrepancy* ($|\Delta P(X)|$) of a rule between two groups. Note that M varies between different epochs (people can have different behavior pattern during the day) as well as rule types (*i.e.*, common or unique). *E.g.*, M of a weekday night rule can be much bigger than that of a weekday morning rule.

No	Rule	*X*	*Y*	Type	*Ctx Spec*	*Conf Diff*	*Cond Disc*	*M*
1	Wkdy Night No.15	- [Location] Avg duration in different frequent places (medium) - [CampusMap] Percentage of time off-campus (low)	- [CampusMap] Percentage of time in sport spaces (low)	Unique Non-dep	2	0.828	0.213	0.695
2	Wkdy Night No.12	- [Sleep] Maximum length of asleep bouts (high) - [Sleep] Sleep Efficiency (low)	- [CampusMap] Percentage of time off-campus (low)	Unique Non-dep	2	0.773	0.232	0.655
3	Wkdy Night No.11	- [Location] Avg duration in different frequent places (medium) - [CampusMap] Percentage of time in sport spaces (low)	- [CampusMap] Percentage of time off-campus (low)	Unique Non-dep	2	0.675	0.261	0.567
4	Wkdy Night No.10	- [Sleep] Maximum length of asleep bouts (high) - [CampusMap] Percentage of time off-campus (low)	- [Sleep] Sleep Efficiency (low)	Unique Non-dep	2	0.616	0.291	0.522
5	Wkdy Night No.9	- [Sleep] Sleep Efficiency (low) - [CampusMap] Percentage of time off-campus (low)	- [Sleep] Maximum length of asleep bouts (high)	Unique Non-dep	2	0.535	0.335	0.453
6	Wkdy Night No.8	- [CampusMap] Percentage of time in sport spaces (low) - [CampusMap] Percentage of time off-campus (low)	- [Location] Avg duration in different frequent places (medium)	Unique Non-dep	2	0.279	0.632	0.234
7	Wkdy Night No.17	- [Sleep] Maximum length of asleep bouts (high)	- [Sleep] Sleep Efficiency (low) - [CampusMap] Percentage of time off-campus (low)	Unique Non-dep	1	0.471	0.38	0.2
8	Wkdy Night No.16	- [Sleep] Number of bouts being awake (low)	- [Sleep] Sleep Efficiency (low)	Unique Non-dep	1	0.418	0.418	0.175
9	Wkdy Night No.13	- [Sleep] Sleep Efficiency (low)	- [Sleep] Maximum length of asleep bouts (high) - [CampusMap] Percentage of time off-campus (low)	Unique Non-dep	1	0.404	0.444	0.171

Table 8. Cont. Table 7 - Top 20 rules from the four epochs of weekdays that capture behavior difference between depression group and non-depression group.

No	Rule	*X*	*Y*	Type	*Ctx Spec*	*Conf Diff*	*Cond Disc*	*M*
10	Wkdy Night No.14	- [Sleep] Sleep Efficiency (low)	- [Sleep] Number of bouts being awake (low)	Unique Non-dep	1	0.394	0.444	0.165
11	Wkdy Afternoon No.1	- [Location] Number of location transition (low) - [Location] Number of on-campus location transition (low) - [CampusMap] Number of building transition on-campus (low) - [CampusMap] Percentage of time in residential spaces (low)	- [Location] Location Variance (low)	Common	4	0.123	0.051	0.039
12	Wkdy Night No.5	- [Bluetooth] Number of unique device of others (low) - [CampusMap] Percentage of time off-campus (low)	- [Sleep] Sleep Efficiency (low)	Common	2	0.137	0.094	0.031
13	Wkdy Afternoon No.6	- [Location] Number of location transition (low) - [CampusMap] Number of building transition on-campus (low) - [CampusMap] Percentage of time in residential spaces (low)	- [Bluetooth] Number of unique device of others (high)	Common	3	0.124	0.052	0.03
14	Wkdy Night No.3	- [Sleep] Sleep Efficiency (low) - [CampusMap] Number of building transition (low) - [CampusMap] Percentage of time in sport spaces (low)	- [CampusMap] Percentage of time off-campus (low)	Common	3	0.103	0.083	0.029
15	Wkdy Afternoon No.3	- [Step] Sum of steps (high) - [CampusMap] Percentage of time in residential spaces (low) - [CampusMap] Percentage of time in sport spaces (low)	- [CampusMap] Percentage of time in academic spaces (medium)	Common	3	0.115	0.055	0.028
16	Wkdy Afternoon No.7	- [Location] Number of location transition (low) - [CampusMap] Number of building transition on-campus (low) - [CampusMap] Percentage of time in residential spaces (low)	- [Location] Location Variance (low) - [Location] Number of on-campus location transition (low)	Common	3	0.117	0.052	0.027

Table 9. Cont. Table 7 - Top 20 rules selected from the four epochs of weekdays that capture behavior difference between depression group and non-depression group. (continued)

No	Rule	*X*	*Y*	Type	*Ctx Spec*	*Conf Diff*	*Cond Disc*	*M*
17	Wkdy Afternoon No.4	- [Location] Number of on-campus location transition (low) - [CampusMap] Number of building transition on-campus (low) - [CampusMap] Percentage of time in residential spaces (low)	- [Location] Location Variance (low) - [Location] Number of location transition (low)	Common	3	0.111	0.056	0.026
18	Wkdy Afternoon No.5	- [Location] Number of on-campus location transition (low) - [CampusMap] Number of building transition on-campus (low) - [CampusMap] Percentage of time in residential spaces (low)	- [Bluetooth] Number of unique device of others (high)	Common	3	0.111	0.056	0.026
19	Wkdy Afternoon No.2	- [Location] Number of location transition (low) - [Location] Number of on-campus location transition (low) - [CampusMap] Number of building transition on-campus (low)	- [Location] Location Variance (low) - [CampusMap] Percentage of time in residential spaces (low)	Common	3	0.104	0.059	0.025
20	Wkdy Morning No.2	- [Sleep] Mean length of awake bouts (low) - [Sleep] Mean length of restless bouts (low) - [Sleep] Maximum length of awake bouts (low)	- [Location] Number of location clusters (low)	Common	3	0.098	0.068	0.024

Table 10. Top 20 rules selected from the four epochs of weekends that capture behavior difference between depression group and non-depression group.

No	Rule	*X*	*Y*	Type	*Ctx Spec*	*Conf Diff*	*Cond Disc*	*M*
1	Wkend Evening No.13	- [Location] Regularity of circadian movement (high) - [Location] Avg duration in different frequent places (low)	- [CampusMap] Percentage of time in social spaces (low)	Unique Non-dep	2	0.948	0.235	0.894
2	Wkend Morning No.13	- [Sleep] Number of bouts being asleep (high) - [Sleep] Number of bouts being restless (high)	- [Sleep] Sleep Efficiency (medium)	Unique (Dep)	2	0.766	0.351	0.793
3	Wkend Afternoon No.11	- [Location] Moving time percentage (high) - [CampusMap] Number of building transition on-campus (low) - [CampusMap] Percentage of time in sport spaces (low)	- [CampusMap] Number of building transition (low)	Unique Non-dep	3	0.626	0.271	0.775
4	Wkend Morning No.15	- [Sleep] Sleep Efficiency (medium) - [Step] Number of active bouts (low)	- [Sleep] Number of bouts being asleep (high)	Unique (Dep)	2	0.646	0.348	0.612
5	Wkend Morning No.16	- [Sleep] Sleep Efficiency (medium) - [Step] Number of active bouts (low)	- [Sleep] Number of bouts being restless (high)	Unique (Dep)	2	0.614	0.348	0.568
6	Wkend Morning No.17	- [Sleep] Sleep Efficiency (medium) - [Step] Number of active bouts (low)	- [Location] Number of location transition (low)	Unique (Dep)	2	0.614	0.348	0.568
7	Wkend Evening No.12	- [Location] Regularity of circadian movement (high) - [CampusMap] Percentage of time in social spaces (low)	- [Location] Avg duration in different frequent places (low)	Unique Non-dep	2	0.585	0.381	0.552

Table 11. Cont. Table 10 - Top 20 rules selected from the four epochs of weekends that capture behavior difference between depression group and non-depression group. (continued)

No	Rule	*X*	*Y*	Type	*Ctx Spec*	*Conf Diff*	*Cond Disc*	*M*
8	Wkend Morning No.14	- [Location] Number of location transition (low) - [Step] Number of active bouts (low)	- [Sleep] Sleep Efficiency (medium)	Unique (Dep)	2	0.582	0.367	0.538
9	Wkend Night No.18	- [Sleep] Mean length of awake bouts (low) - [Step] Avg number of steps during active bouts (low)	- [Sleep] Number of bouts being asleep (low)	Unique (Dep)	2	0.708	0.197	0.529
10	Wkend Evening No.11	- [Location] Avg duration in different frequent places (low) - [CampusMap] Percentage of time in social spaces (low)	- [Location] Regularity of circadian movement (high)	Unique Non-dep	2	0.56	0.398	0.528
11	Wkend Night No.17	- [Sleep] Number of bouts being asleep (low) - [Step] Avg number of steps during active bouts (low)	- [Sleep] Mean length of awake bouts (low)	Unique (Dep)	2	0.68	0.205	0.508
12	Wkend Afternoon No.14	- [Location] Moving time percentage (high) - [CampusMap] Number of building transition on-campus (low)	- [CampusMap] Number of building transition (low) - [CampusMap] Percentage of time in sport spaces (low)	Unique Non-dep	2	0.576	0.295	0.475
13	Wkend Morning No.12	- [Location] Moving time percentage (low) - [Location] Number of location transition (low)	- [Sleep] Sleep Efficiency (medium)	Unique (Dep)	2	0.497	0.425	0.456
14	Wkend Morning No.11	- [Location] Number of location transition (low) - [CampusMap] Number of building transition (low)	- [Sleep] Sleep Efficiency (medium)	Unique (Dep)	2	0.456	0.469	0.422

Table 12. Cont. Table 10 - Top 20 rules selected from the four epochs of weekends that capture behavior difference between depression group and non-depression group. (continued)

No	Rule	*X*	*Y*	Type	*Ctx Spec*	*Conf Diff*	*Cond Disc*	*M*
15	Wkend Morning No.10	- [Location] Number of location transition (low) - [CampusMap] Number of building transition (low)	- [Sleep] Sleep Efficiency (low)	Common	2	0.435	0.483	0.398
16	Wkend Morning No.6	- [Location] Number of location transition (low) - [Sleep] Number of bouts being asleep (high)	- [Sleep] Number of bouts being restless (high)	Common	2	0.095	0.051	0.013
17	Wkend Morning No.5	- [Sleep] Mean length of asleep bouts (medium) - [Sleep] Number of bouts being asleep (high)	- [Sleep] Number of bouts being restless (high)	Common	2	0.082	0.062	0.012
18	Wkend Morning No.2	- [Sleep] Number of bouts being asleep (high) - [Sleep] Number of bouts being restless (high)	- [Sleep] Mean length of asleep bouts (medium)	Common	2	0.071	0.081	0.011
19	Wkend Morning No.3	- [Sleep] Number of bouts being asleep (high) - [Sleep] Number of bouts being restless (high)	- [Location] Number of location transition (low)	Common	2	0.053	0.081	0.007
20	Wkend Morning No.9	- [Screen] Length std of screen having interaction (low) - [Screen] Length std of screen being unlock (low)	- [Screen] Mean length of screen having interaction (low)	Common	2	0.059	0.052	0.006