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.* 37, 4, Article 111 (August 2019), 34 pages. https://doi.org/10.1145/1122445.1122456

Authors' addresses: Xuhai Xu, 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.

2474-9567/2019/8-ART111 $15.00
https://doi.org/10.1145/1122445.1122456

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.

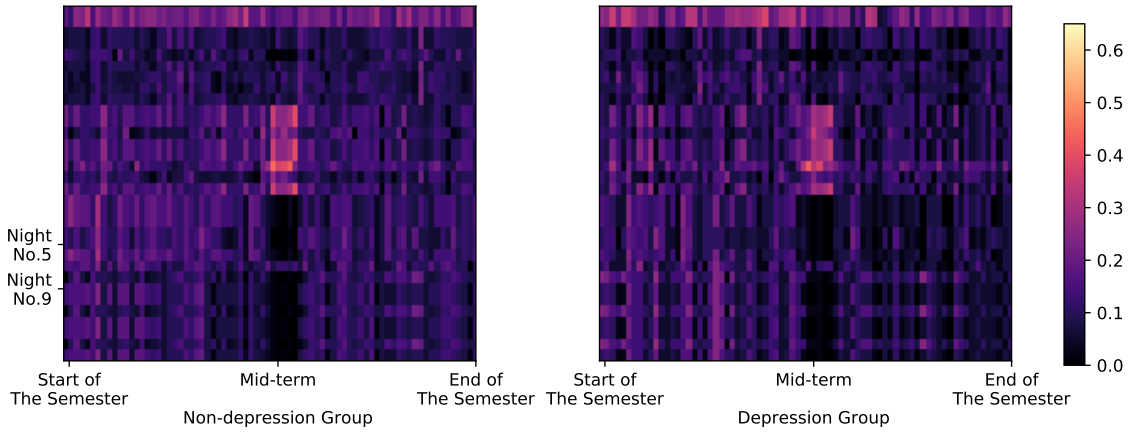
4.3.3 Feature Preparation before Rule Mining. ARM is typically applied on symbolic or categorical data. We therefore recoded each of the selected features, for rule selection using ARM only, into the three categories: low, moderate, and high, using a binning method. Each category contained 33.3% of each 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.

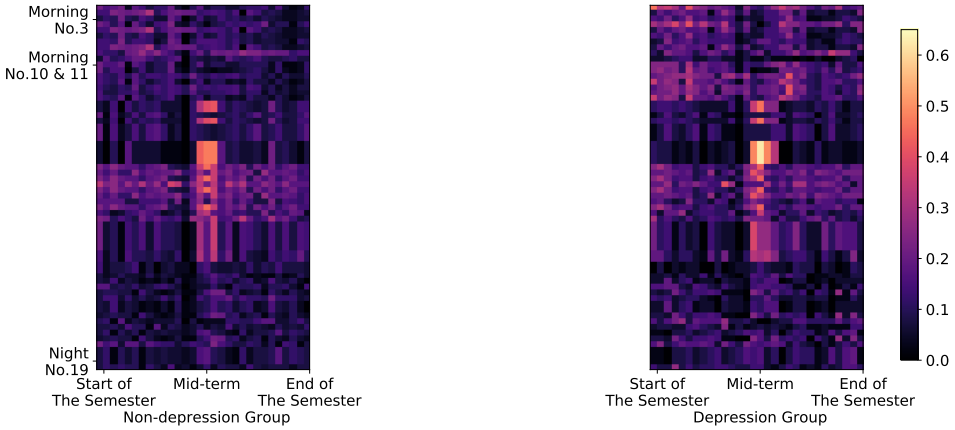
4.3.5 Feature Extraction and Model Training. After we obtained the rules, we turned to the TrainTestSet and used Algorithm 2 to extract an average of 17 contextually filtered features (Min = 8, Max = 23) per epoch, 137 in total. Note that one rule $X \rightarrow Y$ can have multiple features y in Y, 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.



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



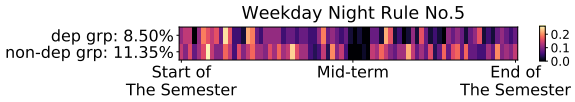
(b) Weekend Rules (Left: Students with no depressive symptoms, Right: depressive symptoms)

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).

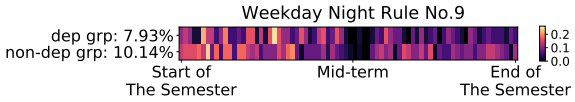
5.1 Rules Can Capture Routines Behaviors and Behavior Pattern Differences between Groups

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.

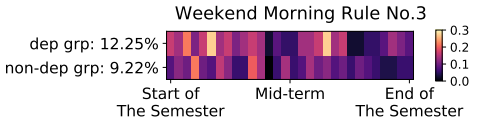
Figure 3 visualizes heatmaps that represent how many students' behavior was captured by each of the 105 rules throughout the study period. From both heatmaps of weekdays and weekends, we observed abnormal color patterns during the middle of the study. The academic calendar of the university showed that this period was



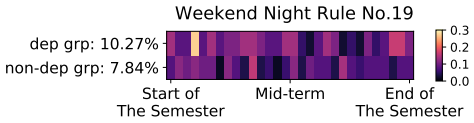
(a) A Rule Common in Groups. More in Non-dep Group



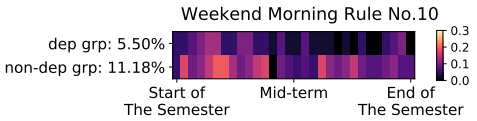
(b) A Rule that is Unique to Non-dep Group



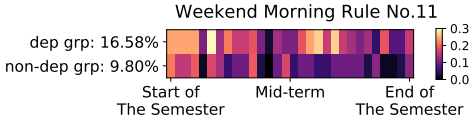
(c) A Rule Common in Groups. More in Dep Group



(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.

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 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).

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.

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 $p < 0.001$

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.

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).

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

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%***

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 ($p < 0.01$).

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).

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

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

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.

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* ($w_3 = 0.5$).

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.

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%***

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 ($p < 0.001$).

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 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.

9 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, University of Washington College of Engineering, Samsung, 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

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 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*
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 7. Top 20 rules 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 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*
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 8. 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*
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 8. 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

Table 8. Top 20 rules selected from the four epochs of weekends that capture behavior difference between depression group and non-depression group. (continued)