Modeling Wellness using Smartphone and Activity Data

Ather Sharif
Paul G. Allen School
University of Washington
ather@cs.washington.edu

Galen Weld
Paul G. Allen School
University of Washington
gweld@cs.washington.edu

Xuhai Xu
Information School
University of Washington
xuhaixu@cs.washington.edu

Abstract

The proliferation of smartphones and the associated "always-connected" lifestyle has raised concerns about their users' wellness. Amongst college students, rates of self-reported mental health challenges are increasing across the vast majority of universities in the United States. In addition to their potential negative effects on wellness, mobile phones offer an unprecedented ability to monitor students' day-to-day and sleep habits. We propose a new method to monitor students' mental health using data from smartphones and physical activity, with a focus on depression. We improve upon previous depression-detection work by demonstrating a longitudinally generalizable model using our new method. Our generalizable model significantly outperforms the baseline model by 6.3% on the balanced accuracy and 0.121 on the κ value.

1 Introduction

Rates of mental health disorders appear to be increasing on college campuses across the United States [1]. Some studies have connected this increase with a simultaneous increase in the use of social media services such as Facebook and Instagram, enabled by the rapid proliferation of smartphones and widespread internet access [2]. However, the ubiquity of smartphones also results in far higher-resolution monitoring of activity such as location, sleep, physical activity, and phone/app usage than has been possible in the past. This great access to personal data is not without significant privacy concerns [3]. Preliminary work by [4, 5, 6] finds that the use of these features offers significant potential to accurately predict depression in college students. For instance, [6] achieve 81.3% balanced accuracy using a contextually-filtered feature extraction system based upon association rule mining.

However, these numbers representing accuracy suffer from a significant caveat – the performance of the models does not generalize beyond the single semester-long study period for which data was recorded. For the potential benefit of this work to materialize, good generalizability is critical, as a

Preprint. Under review.

system must perform well on subjects over a period of time longer than a 16-week semester. There are important research questions related to the longitudinal generalizability for depression sensing: **1) can we leverage data collected earlier to train a model and apply it on the same users to monitor their depression status in the future?** and **2) can we train a model with our current data from a group of users, and then use this model to recognize other users' depressive symptoms in the future?**

To achieve these goals, we demonstrate a new approach combining several new feature selection techniques, which achieves significantly better performance than the previous work [6] when generalized to different semesters than it was trained upon. To the best of our knowledge, we are the first to address the longitudinal generalizability on the model level in the depression sensing area, with a dataset over two discontinuous academic semesters (spring 2017 and spring 2018).

Specifically, we propose a three-stage method to train a longitudinally generalizable model, including feature selection, feature filtering, and feature purification (see Figure 1). The three steps can greatly simplify the features used during model training and produce a model that significantly outperformed the baseline model by 6.3% on the balanced accuracy and 0.121 on the κ value.

We review relevant prior work in Section 2, and introduce our method in Section 4. Then, we highlight the uniqueness of our data sources together with the research questions in Section 3, followed by the details of the method implementation in Section 5. We present our experiment results in Section 6. Finally we discuss and summarize our findings in Section 7 and Section 8.

2 Related Work

2.1 Mobile Sensing and Depression Detection

In recent years there has been a dramatic increase in the volume of research focused on mobile sensing for wellness. Wang et al. [4] at Dartmouth studied stress in 48 college students across a 10 week term, finding correlations between phone sensor data and stress levels, as well as a 'lifecycle' of stress aligned with the academic calendar, where stress levels are low at the beginning of the term, and increase with midterm and final exams. A follow-up study [5] also looked at using the same data to predict academic performance (GPA), which is known to be correlated with mental health [7].

Mobile sensors have also been used to detect depression, and past work has further aided in this detection by incorporating machine learning models. Saeb et al. [8] 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 identifying between participants with and without depressive symptoms. Similarly, Canzian and Musolesi [9] collected data from another group of 28 participants, and trained models on an individual level with location features to detect periods when users experienced depression. However, previous work relied on a single dataset and none of them discussed the generalizability of their models.

2.2 Contextually-Filtered Features For Depression Detection

Xu et al. [6] collected data from college students over two discontinuous academic semesters. The first dataset contained 138 college students collected in the spring of 2017, and the second dataset contained 212 students collected in the spring of 2018. They proposed a new pipeline to extract what they defined as contextually filter features for depression detection. First, the authors extracted a feature pool of more than 200 features based on mobile sensing data. After selecting the top features based on mutual information, they used association rule mining [10] to extract the rules of the co-occurrence of the sensor features. Then, they devised new metrics to pick up the top rules that could best capture the difference between the student group that had depressive symptoms and the group that did not. They proposed an algorithm to extract contextually features based on the selected rules. Finally, they train a classifier on the two student groups based on the contextually features. The results showed that the new features could achieve better performance on various metrics such as accuracy and F1 score.

Furthermore, they also tested the generalizability of their method across the two datasets. They showed the generalizability of their pipeline (*i.e.*, repeated the whole pipeline on the second dataset) and the rules (*i.e.*, obtained the association rules from the first dataset and did the training on the

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

Table 1: Sensor data and information aggregated into features.

Study	Days	Dataset	Post-semester BDI-II		Overlap	Post-semester BDI-II	
		Size	Non-dep Grp	Dep Grp	Size	Non-dep Grp	Dep Grp
Phase I	106	137	81	56 (40.9%)	68	44	24 (35.3%)
Phase II	113	218	139	79 (36.2%)	68	41	27 (39.7%)

Table 2: Information of the two studies. Students with a post-semester Beck Depression Inventory score greater than 13 were in the depression group, in accord with the interpretation of the BDI-II [12].

second dataset), but not on the model level (*i.e.*, obtained the rules and trained the model on the first dataset and applied it on the second dataset directly). In this paper, we introduce a new simplified pipeline to obtain a model that is generalizable across the two datasets.

3 Data Sources

We leveraged a dataset collected by one of our collaborative research groups to verify our method. The dataset has two phases, including data from smartphone activity and mental health for college students, collected over two disjoint semester-long periods in 2017 and 2018. 138 participants' data were collected in the Phase I dataset (2017 Spring), whereas 212 participants' data were collected in the Phase II dataset (2018 Spring). Among the two datasets, there is an overlap of 68 students whose data were collected in both phases.

We make use of these two phases as they provide a lens to explore whether we can train on the data from one semester to predict students' depression status in another semester, critical to our two research questions from Section 1: **RQ 1)** Can we leverage the data that is collected earlier to train a classification model and apply it on the same users to monitor their depression status in the future? **RQ 2)** Can we train a model with our current data from a group of users, and apply it on other users to recognize their depressive symptoms in the future?

To answer RQ1, we focus on the 68 overlapping students to investigate generalizable models. In addition, to address RQ2, we also implement our method on all participants from the Phase I dataset, and apply it on the Phase II dataset (with the overlapping students excluded from the Phase II dataset, *i.e.*, non-overlapping students), and test the results.

The data covers a broad range of modalities: a baseline survey collected information about academic history, lifestyle, and demographic information. For the entire study period, sleep and physical activity information were collected using a Fitbit Flex 2. Phone usage information, including Bluetooth, location, call, and screen-time were collected using the AWARE framework [11]. The total size of the raw data is close to 300 gigabytes. Table 1 summarizes the sensors, their corresponding raw features, as well as the average number of the samples per person. Each feature was calculated on daily-episode basis, with the day divided into four episodes: night (12am-6am), morning (6am-12pm), afternoon (12pm-6pm), and evening (6pm-12am). Overall, 212 daily-episode features were extracted from the raw sensor streams.

The dataset employs Beck Depression Inventory-II (BDI-II) [12], a widely used psychometric test for depressive symptoms severity measurement, as the ground truth label of whether students experienced depressive symptoms at the end of the semester. A student was labeled as having depression at the end

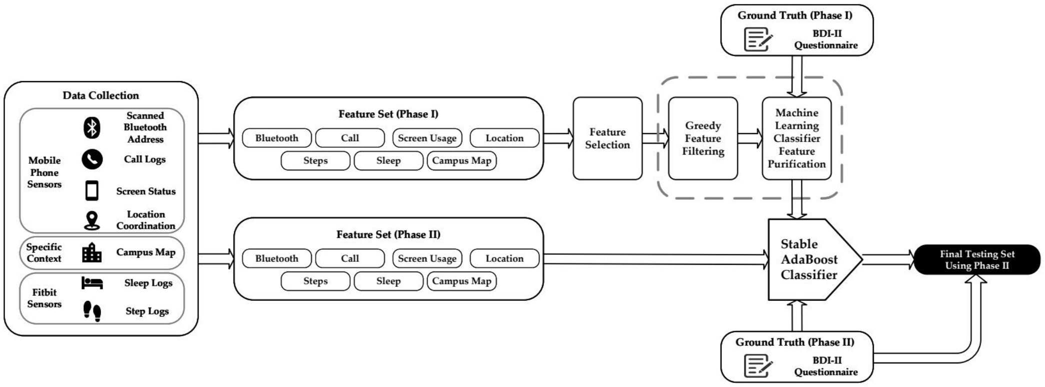


Figure 1: The high-level pipeline of our method. The dashed frame highlights the novel contribution of the paper. We design two new parts to select the features for training a longitudinally generalizable model. Note that the Phase I dataset and Phase II dataset refers to two datasets collected from the same group of people during two different longitudinal periods.

of the semester (*i.e.*, the depression group) if the BDI-II score was higher than 13 [13], otherwise, they were placed in the non-depression group. The ground truth information is summarized in Table 2.

4 Methods

Frequent sampling of phone usage and activity results in an enormous quantity of data — our raw dataset for both phases is approximately 300 gigabytes. As such, finding the most important features from this rich dataset is the key to successfully training a generalizable classification model. In this section, we introduce our method of picking up the best sensor features to train a model that is generalizable longitudinally, *i.e.*, a model trained on an earlier dataset can be applied directly on a later dataset. Our method emphasizes significantly the feature selection process and combines three stages to obtain a stable mobile sensing feature set. Figure 1 visualizes the overall pipeline.

4.1 Feature Selection

Using all the features from the rich mobile sensing data for training can easily lead to over-fitting, as the number of features is greater than the number of data points. Feature selection is a common step employed by previous works [8, 14]. We use mutual information [15] to select features. Specifically, we use the method proposed by Kraskov et al. [16] to estimate the mutual information gain and pick the top k features. Because of the random noise introduced to avoid duplicates during the calculation, each time we obtain a different set of top features. Therefore, we calculate the top feature set repeatedly and select the overlap features iteratively until they converge.

4.2 Feature Filtering

However, mutual information does not take correlation among features into account. There can be highly correlated features among the top features selected from Section 4.1, for example, `heart rate` and `activity level`. To tackle this, we devise a loop-based greedy algorithm to remove highly correlated features after the mutual-information-based feature selection. In each loop, we pick the pair of features that had the highest Pearson correlation coefficient. For each of the two features in the pair, we calculated the sum of the correlation coefficients with all other remaining top features. Then, we remove the one that had a higher sum from the top feature set. We repeat the loop until all correlation coefficients among the remaining features are less than a pre-determined threshold of α.

4.3 Feature Purification during Training

After feature selection and filtering, we obtain a subset of the raw features that can be used for training. We have selected features that have a high dependency with the ground truth label (from

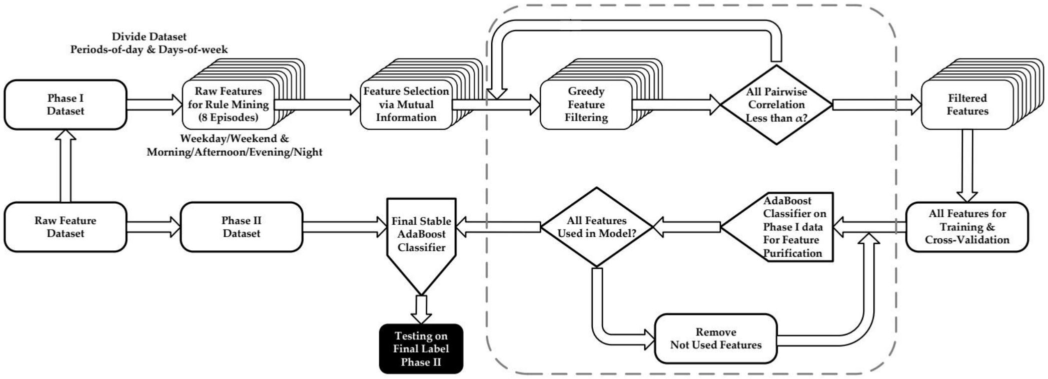

Divide Dataset
Periods-of-day & Days-of-week
Phase I Dataset
Raw Features for Rule Mining (8 Episodes)
Weekday/Weekend & Morning/Afternoon/Evening/Night
Feature Selection via Mutual Information
Greedy Feature Filtering
All Pairwise Correlation Less than α?
Filtered Features
All Features for Training & Cross-Validation
AdaBoost Classifier on Phase I data For Feature Purification
All Features Used in Model?
Remove Not Used Features
Final Stable AdaBoost Classifier
Testing on Final Label Phase II
Raw Feature Dataset
Phase II Dataset

Figure 2: The detailed pipeline of feature selection, filtering and purification. The dashed frame highlights the novel procedures in the pipeline.

the definition of mutual information), and that are not strongly correlated with each other (from the filtering step). However, the intrinsic relationships among the features, as well as the convoluted effects of combining these features when classifying the final label, still remain implicit. Some machine learning models such as a decision tree can perform feature selection during the training, which to an extent captures the relationship amongst and between the features and the labels through some metrics (*e.g.*, information gain when training C4.5 decision tree [17]). We leverage this property of decision trees to eliminate the unnecessary features that otherwise may introduce noise and bias during the training and purify for those important features.

We employ AdaBoost [18] with decision-tree-based component classifiers as the model for the classification task. To maintain the generalizability, we keep each decision-tree component to be shallow and thus, make use of a further smaller fraction of the filtered feature set. To purify for these important features, we perform a recursive purification by only retaining the features that are used in the last run of the training. The loop ends until all remained features are used by the model, which means that the feature set is stabilized.

Figure 2 shows the details of the pipeline. Through these three steps, we obtain a feature set that is significantly smaller than the raw feature set. We then use this feature set to train a model from an anterior dataset and test on a posterior dataset.

5 Implementation of the Method

We introduce how we implemented our method on our dataset. In both phases of the dataset, the data from each participant covers an entire 16 week semester. We calculated the mean and standard deviation of all daily-episode feature across all weekdays and weekends individually to aggregate the features. In an approach similar to [6], we split the week to capture different behavior patterns on weekdays and weekends, as suggested by past literature [19], resulting in 8 episode groups (4 daily episodes × 2 types of days-of-week). This gave 424 aggregated daily-episode features per participant (212 × 2). Our feature selection and filtering steps were applied to each episode group respectively. The feature purification step was applied after concatenating the features. Note that all these steps were applied to the anterior training dataset, *i.e.*, the phase I dataset. The Phase II dataset was only used for testing. The details of the implementation results are not exactly the same when implementing our method with different focuses (*i.e.*, either the overlapping students for RQ1 or the non-overlapping students between the two phases for RQ2), here we present the implementation with the focus on the 68 overlapping students.

Feature selection Using all 3392 daily-episode features (424 aggregated features × 8 episode-group) for training on a dataset with 68 data points is not feasible. We performed the feature selection described in Section 4.1 with k set to 100. In each iteration, we picked the top 100 features, and retain the overlapping features. The feature set converged after 30 iterations. We obtained 42 features on average (Min = 32, Max = 52, 335 in total across the 8 episode groups).

Feature filtering We then iteratively removed the features one by one using the algorithm introduced in Section 4.2. We set the upper correlation coefficient threshold α as 0.5, giving an average of 19 features (Min = 15, Max = 21, 148 in total across the 8 episode groups).

Feature purification After concatenating features across episodes, we embedded the purification step into the training and hyper-parameter tuning step. Any time a model was trained, the purification would take place and ensure the features used for training were stabilized. We employed the Leave-One-Person-Out (LOPO) cross-validation technique as it is widely used in the mobile sensing area [8, 14]. Parameter tuning was based on the average balanced accuracy of the leave-one-out validation results. We performed a grid search over the two hyper-parameters of the AdaBoost model: the number of estimators (search range: $3, 5, 7, 10, 15, 20$) and either the maximum depth (search range: $2, 3, 4, 5$) or the maximum number of leaf nodes (search range: $3, 4, 5, 6, 7, 8, 9$) of each decision-tree component. We chose the number of estimators to be 5, and the maximum number of leaf nodes to be 6. The stabilized feature set contained 11 features.

Training and testing Finally, we used these features to train a decision tree model on the Phase I dataset and tested it on the Phase II dataset.

6 Results

We present the performance of the classification model based on our new method. We first report the results for RQ1 in Section 6.1. Our method generates a model that outperforms the best baseline model by 6.3% on the absolute balanced accuracy and 0.121 on the κ value. In addition, we also show the results addressing RQ2 in Section 6.2. Our model achieves an increase of 2.7% on the balanced accuracy and 0.055 on the κ value over the best baseline model.

6.1 Longitudinal Generalizability on Overlapping Students

To answer RQ1, we set the focus on the 68 overlapping students. We compared our new model with a few baselines: 1) majority, simply categorizing all samples as the majority type in the dataset, which was non-depression; 2) single feature, using the one best features to do the classification; 3) full features, using all sensors' features for the classification; 4) a common practice approach using features after direct selection. Feature selection based on certain ranking metrics is common practice in previous works, *i.e.*, selecting the top features ranked by metrics. We employed mutual information (*e.g.*, [20]) as the metrics for this baseline, a similar process as the feature selection introduced in Section 4.1. Thus, one can imagine the common practice baseline as only applying the first step of our method.

Moreover, in order to better understand the effectiveness of each of the three steps (feature selection, filtering, and purification) in our method, we further performed a "step-ablation" study: removing each of the step at one time and compare the results.

Table 3 summarizes the results of various metrics. Our model (the last row of Table 3) achieves the highest scores on the balanced accuracy and the κ value. Compared to the best baseline model, trained on the features after selection (only stage 1), our model has a significant increase of 6.3% on the balanced accuracy, as well as 0.121 on the κ value ($p < 0.01$).

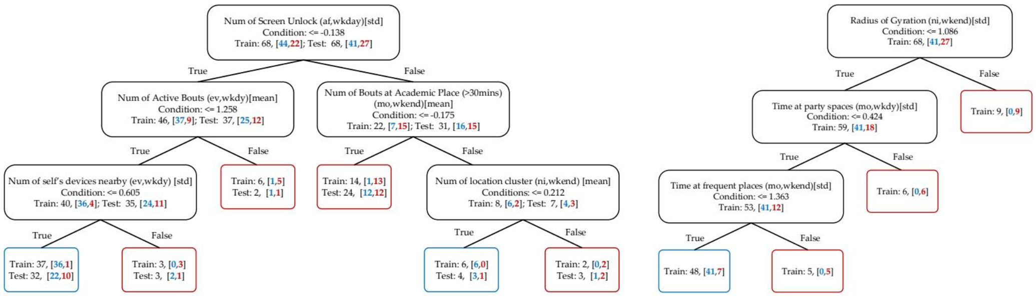
Comparing among the models from the "step-ablation" study, the one without step 1 has the greatest decrease of the balanced accuracy, and the one without step 2 has the least decrease over the two metrics. The results reflect the relative importance of the three-step to some extent: feature selection is the most important, followed by feature purification, followed by feature filtering.

We visualize one of the decision trees in our model to obtain a better interpretation of the classifier, as well as a better understanding of the students' behavior, shown in Figure 3a. This is discussed further in Section 7.

We compared the classification results of students in two phases (the "train" group versus the "test group"). A larger proportion of the students in the non-depression group spent little time in academic places during Phase II (12 out of 41) compared to Phase I (1 out of 44), which was one of the major misclassification cases in the testing set. It was a common phenomenon that students could spend less time in academic places as their grade proceeded.

Method	Features	Sensitivity	Specificity	Balanced Accuracy	κ
Baseline: Majority	-	0.000	1.000	0.500	0.000
Baseline: Single Feature	Best Single Feature	0.518	0.536	0.527	0.053
Baseline: Full Features	All Features	0.481	0.585	0.533	0.066
Baseline: Common Practice	Features after selection (Stage 1)	0.417	0.659	0.538	0.075
New method without Step 1	Features after filtering&purification	**0.704**	0.390	0.547	0.085
New method without Step 2	Features after selection&purification	0.592	0.585	0.589	0.171
New method without Step 3	Features after selection and filtering	0.592	0.536	0.564	0.123
New method with All Three Steps	Feature after selection, filtering, purification	0.593	**0.610**	**0.601**	**0.196**

Table 3: Comparison of the test results of the baseline classifiers and our new method. The models above the dashed line are baselines. All models are AdaBoost [18] with decision-tree-based component classifiers (number of estimators is 5 and the maximum number of leaf nodes is 6). Note that these models are trained on the overlapping students in Phase I dataset and tested on the same students in Phase II dataset.



(a) Tree trained on Phase I and tested on Phase II

(b) Tree trained on Phase II

Figure 3: Visualization of One of The Final AdaBoost Classifier's Decision Trees. (a) Leave-one-person-out (LOPO) cross-validation (on Phase I) balanced accuracy is 0.837, the testing accuracy (on Phase II) is 0.601. (b) As an illustration of the difference between two phases, LOPO cross-validation (on Phase II) balanced accuracy is 0.748, with the tree significantly different from (a). In each node of the trees, the "train" indicates the size of the two group types in the training set, while the "test" indicates the size in the testing set. The blue color indicates the non-depression group and the red color indicates the depression group. The border color of the leaf nodes imply the final classification labels of the samples in the leaf nodes.

In addition, we trained another model on the dataset from Phase II to identify the major features to recognize students in the depression group of Phase II. Figure 3b visualizes one of the decision tree, whose nodes used the standard deviations of the features related to location for the decision function, indicating that students in the depression group had greater location variance compared to students in the non-depression group.

Method	Features	Sensitivity	Specificity	Balanced Accuracy	κ
Baseline: Majority	-	0.000	1.000	0.500	0.000
Baseline: Single Feature	Best Single Feature	0.161	0.901	0.531	0.069
Baseline: Full Features	All Features	0.339	0.753	0.546	0.097
Baseline: Common Practice	Features after selection (Stage 1)	0.232	0.827	0.530	0.064
New method without Step 1	Features after filtering&purification	0.232	**0.827**	0.530	0.064
New method without Step 2	Features after selection&purification	**0.438**	0.578	0.508	0.015
New method without Step 3	Features after selection and filtering	0.333	0.618	0.475	-0.047
New method with All Three Steps	Feature after selection, filtering, purification	0.393	0.753	**0.573**	**0.152**

Table 4: Comparison of the test results of the baseline classifiers and our new method. The models above the dashed line are baselines. All models are using AdaBoost [18] with decision-tree-based component classifiers (number of estimators is 5 and the maximum number of leaf nodes is 6).

6.2 Longitudinal Generalizability on The Non-overlapping Students

To answer RQ2, we switched the focus to the non-overlapping students in two phases. We excluded the 68 overlapping students in Phase II, leading to a dataset of 150 students (218 - 68). We then applied our method on the Phase I dataset (138 students) to train a model and tested it on the Phase II dataset (150 different students). We conducted the same experiment and the results are summarized in Table 4.

Our method produced a model that outperforms the best baseline model, the one trained on full features, by 2.7% on the balanced accuracy, as well as 0.055 on the κ value (marginal significance $p = 0.08$). Compared to the results in Table 3, our model on the non-overlapping students has a smaller advantage over the baseline models. This indicates greater difficulties of generalizing a model across different users than a model among the same group.

7 Discussion

The results presented in Section 6 present a preliminary demonstration that our new method significantly outperforms current best practices when generalized across both times (train on Phase 1, test on Phase 2) and users (train on overlapping Phase 1 users, test on non-overlapping Phase 2 users).

Although other methods achieve higher immediate accuracy (*e.g.* accuracy computed on a test set taken from the same population at the same time period as the train set), the generalizability of our model is superior. For example, [6] found a 54.2% accuracy when applying their model trained upon Phase 1 data to Phase 2 data – barely better than 50/50 chance.

When examining Figure 3, which displays the decision trees produced by our training model, we see that our decision tree echoes many common characteristics of depressed people found in the psychology literature:

- A large proportion of students in the depression group (15 out of 22) had a high variance in using smartphones in weekday afternoons. This is in line with the depressive symptoms of *diminished ability of concentration* [21]. Students could be more easily distracted by smartphones, leading to a higher variance of phone interaction times.

- A majority of students in the depression group spent less time on academic places on weekend mornings, showing a less interest in the study, which was also a reflection of the symptom of *diminished ability of concentration*, as well as the symptom of *loss of energy* [21, 22].
- However, some students in the depression group had higher activity bouts in weekday evenings, which indicates more physical activities, and more location clusters in weekend nights, which indicates higher mobility variance. These were different from the symptoms of *diminished interests or pleasure in activities* [21]. One explanation might be that the decreased ability of concentration and interests led to frequent switches between various activities.

7.1 Limitations and Future Work

Although our work presents a promising start, there are significant possibilities for future enhancements as well as areas for further exploration.

One such possibility revolves around the nature of the time-series data recorded for each participant. Under the current feature-generation paradigm, these features are binned by a fixed time interval, for example, `Time spent at party spaces during the morning quarter of each weekday`. This simple strategy is effective, but performance could be improved by experimenting with more sophisticated interpretations of these time series data. It's not difficult to imagine that, for the purposes of predicting depression, significantly longer or shorter time-bins would be critical to the identification of activity patterns predictive of depression. One possibility would be to develop some more complex models of activity over time, and incorporate these directly into the model, as an alternative to binning.

Given more time, we would also like to experiment more with different ranges of depth and number of leaf nodes for our trees. The possibility space in this regard is enormous. Although a significant increase in depth diminishes both the interpretability of the model, as well as opens up greater potential for overfitting, we believe that we could further improve our performance with more experimentation.

Lastly, as is the case with essentially every machine learning system, our performance is limited by the volume of data available. Although we have a large amount of data for each participant, our number of participants, especially the 68 participants who participated in both phase 1 and phase 2, is constraining. People are diverse, and no one person is representative of all people experiencing depression. We overcome this challenge with volume, and a greater volume of participants would further improve the quality of our model as well as our confidence in our results.

8 Conclusion

Mental health is a critical issue, and we believe our work demonstrates the significant potential that rich sensing data provided by smartphones has to improve early detection of depression and other mental health challenges. Early recognition of deteriorating mental health is key to successful treatment and minimizing the impacts to a patient's life, and, in the case of students, their academic career.

Most work on depression detection has focused on a single semester worth of data. We, however, feel strongly that this short time period is not adequate for impact in the real world. For phone-based mental health systems to be truly useful, they must work well over a long time period, and they must work for multiple users – they must generalize both across time, and across users. In this paper, we focus on this important aspect of detection depression by leveraging a dataset that spans more than a year. We look forward to seeing further improvement in this important and exciting research space.

9 Acknowledgements

We would like to thank Jennifer Mankoff for providing us with guidance and access to this sensitive dataset. We would also like to thank Alex Okeson for her contributions and valuable advice as our go-to TA for this project.

References

[1] Justin Hunt and Daniel Eisenberg. Mental health problems and help-seeking behavior among college students. *Journal of adolescent health*, 46(1):3–10, 2010.

[2] Igor Pantic. Online social networking and mental health. *Cyberpsychology, Behavior, and Social Networking*, 17(10):652–657, 2014.

[3] Janice C Sipior, Burke T Ward, and Linda Volonino. Privacy concerns associated with smartphone use. *Journal of internet commerce*, 13(3-4):177–193, 2014.

[4] Rui Wang, Fanglin Chen, Zhenyu Chen, Tianxing Li, Gabriella Harari, Stefanie Tignor, Xia Zhou, Dror Ben-Zeev, and Andrew T. Campbell. 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*, UbiComp '14, pages 3–14, New York, NY, USA, 2014. ACM.

[5] Rui Wang, Gabriella Harari, Peilin Hao, Xia Zhou, and Andrew T. Campbell. Smartgpa: How smartphones can assess and predict academic performance of college students. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '15, pages 295–306, New York, NY, USA, 2015. ACM.

[6] Xuhai Xu, Prerna Chikersal, Afsaneh Doryab, Daniella Villalba, Janine Dutcher, Michael Tumminia, Tim Althoff, Sheldon Cohen, Kasey Creswell, David Creswell, Jennifer Mankoff, and Anind K. Dey. Leveraging routine behavior and contextually-filtered features for depression detection among college students. In *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*. Please contact the team for this paper that is under major revision.

[7] Michelle Richardson, Charles Abraham, and Rod Bond. Psychological correlates of university students' academic performance: A systematic review and meta-analysis. *Psychological bulletin*, 138(2):353, 2012.

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

[9] Luca Canzian and Mirco Musolesi. 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*, pages 1293–1304. ACM, 2015.

[10] Rakesh Agrawal, Ramakrishnan Srikant, et al. Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB*, volume 1215, pages 487–499, 1994.

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

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

[13] David JA Dozois, Keith S Dobson, and Jamie L Ahnberg. A psychometric evaluation of the beck depression inventory–ii. *Psychological assessment*, 10(2):83, 1998.

[14] Fabian Wahle, Tobias Kowatsch, Elgar Fleisch, Michael Rufer, and Steffi Weidt. Mobile Sensing and Support for People With Depression: A Pilot Trial in the Wild. *JMIR mHealth and uHealth*, 4(3):e111, 2016.

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

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

[17] J Ross Quinlan. *C4. 5: programs for machine learning*. Elsevier, 2014.

[18] Yoav Freund and Robert E Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139, 1997.

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

[20] Sangwon Bae, Denzil Ferreira, Brian Suffoletto, Juan C Puyana, Ryan Kurtz, Tammy Chung, and Anind K Dey. Detecting drinking episodes in young adults using smartphone-based sensors. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(2):5, 2017.

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

[22] Rui Wang, Weichen Wang, Alex daSilva, Jeremy F. Huckins, William M. Kelley, Todd F. Heatherton, and Andrew T. Campbell. Tracking depression dynamics in college students using mobile phone and wearable sensing. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 2(1):43:1–43:26, March 2018.