Leveraging Collaborative-Filtering for Personalized Behavior Modeling: A Case Study of Depression Detection among College Students

XUHAI XU, University of Washington
PRERNA CHIKERSAL and JANINE M. DUTCHER, Carnegie Mellon University
YASAMAN S. SEFIDGAR, University of Washington
WOOSUK SEO, University of Michigan
MICHAEL J. TUMMINIA, University of Pittsburgh
DANIELLA K. VILLALBA, SHELDON COHEN, KASEY G. CRESWELL, and J. DAVID CRESWELL, Carnegie Mellon University
AFSANEH DORYAB, University of Virginia
PAULA S. NURIUS, EVE RISKIN, ANIND K. DEY, and JENNIFER MANKOFF, University of Washington

The prevalence of mobile phones and wearable devices enables the passive capturing and modeling of human behavior at an unprecedented resolution and scale. Past research has demonstrated the capability of mobile sensing to model aspects of physical health, mental health, education, and work performance, *etc.* However, most of the algorithms and models proposed in previous work follow a one-size-fits-all (*i.e.*, population modeling) approach that looks for common behaviors amongst all users, disregarding the fact that individuals can behave very differently, resulting in reduced model performance. Further, black-box models are often used that do not allow for interpretability and human behavior understanding. We present a new method to address the problems of personalized behavior classification and interpretability, and apply it to depression detection among college students. Inspired by the idea of collaborative-filtering, our method is a type of memory-based learning algorithm. It leverages the relevance of mobile-sensed behavior features among individuals to calculate personalized relevance weights, which are used to impute missing data and select features according to a specific modeling goal (*e.g.*, whether the student has depressive symptoms) in different time epochs, *i.e.*, times of the day and days of the week. It then compiles features from epochs using majority voting to obtain the final prediction. We apply our algorithm on a depression detection dataset collected from first-year college students with low data-missing rates and show that our method outperforms the state-of-the-art machine learning model by 5.1% in accuracy and 5.5% in F1 score. We further verify the pipeline-level generalizability of our approach by achieving similar results on a second dataset, with an average improvement of 3.4% across performance metrics. Beyond achieving better classification performance, our novel approach is further able to generate personalized interpretations of the models for each individual. These interpretations are supported by existing depression-related literature and can potentially inspire automated and personalized depression intervention design in the future.

Authors' addresses: Xuhai Xu, xuhaixu@uw.edu, University of Washington, 1410 NE Campus Parkway, Seattle, WA; Prerna Chikersal; Janine M. Dutcher, Carnegie Mellon University, Pittsburgh, PA; Yasaman S. Sefidgar, University of Washington; Woosuk Seo, University of Michigan, Ann Arbor, MI; Michael J. Tumminia, University of Pittsburgh, Pittsburgh, PA; Daniella K. Villalba; Sheldon Cohen; Kasey G. Creswell; J. David Creswell, Carnegie Mellon University; Afsaneh Doryab, University of Virginia, Charlottesville, VA; Paula S. Nurius; Eve Riskin; Anind K. Dey; Jennifer Mankoff, University of Washington.

2474-9567/2021/3-ART42 $15.00
https://doi.org/10.1145/3448107

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

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

ACM Reference Format:
Xuhai Xu, Prerna Chikersal, Janine M. Dutcher, Yasaman S. Sefidgar, Woosuk Seo, Michael J. Tumminia, Daniella K. Villalba, Sheldon Cohen, Kasey G. Creswell, J. David Creswell, Afsaneh Doryab, Paula S. Nurius, Eve Riskin, Anind K. Dey, and Jennifer Mankoff. 2021. Leveraging Collaborative-Filtering for Personalized Behavior Modeling: A Case Study of Depression Detection among College Students. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 5, 1, Article 42 (March 2021), 27 pages. https://doi.org/10.1145/3448107

1 INTRODUCTION

As close daily companions to humans, mobile phones and wearable devices can passively capture various aspects of daily routine behavior. A large body of work has demonstrated the feasibility of mobile sensing in many domains, such as monitoring physical health status [8, 39], detecting mental health problems [59, 62], tracking education flow [4], evaluating work performance [40], and promoting social justice [50]. Researchers have employed several methods to tackle their research questions, from inspecting statistical relationships to building machine learning models for certain tasks. However, most of the prior work on algorithms and models follows a one-size-fits-all (population modeling) approach, *e.g.*, building one model for all users to detect depression [60, 62]. Such an approach de-emphasizes the important fact that individuals behave differently. A person may behave similarly to one group but very differently from another group (*e.g.*, [14, 41]). Even within the same subpopulation, no two individuals share identical behaviors. Applying one model to all users does not effectively recognize or leverage such differences, with reduced model performance and interpretability. Moreover, inspecting a model that treats all users as a whole can only reveal common behaviors among the population. Thus it does not provide personalized interpretability, which is an important factor for behavior understanding and modeling. Therefore, there is a growing consensus that a personalized model will perform better than a population model [28].

However, it is not straightforward to just create personalized models instead of population models. Training a personalized model usually requires a large amount of data (especially ground truth labels) from each person to establish a good individual profile. In some domains, such as activity recognition [26, 34], labels are easier to obtain, as opposed to domains like mental health monitoring. For many passive mobile sensing studies, obtaining ground truth is expensive. For example, to get reliable labels of whether a person is experiencing depressive symptoms, users need to complete a well-established survey (*e.g.*, PHQ-9 [31] or BDI-II [9]). Filling in the survey is time- and energy-consuming, which can easily lead participants to drop out of the study, especially when a study is longitudinal and lasts several months or even years. There are a few efforts trying to mitigate this issue. For example, PHQ-4 [32] is designed to be a very short screening scale for anxiety and depression. However, it can only be reliably administered every two weeks, resulting in sparse ground truth [60]. There is often a trade-off between the survey frequency (obtaining ground truth) and participants' compliance or desire to continue their participation (collecting behavior data). Many longitudinal studies involving passive mobile sensing end up with a large amount of mobile sensing data, but very few labels (*e.g.*, [18, 38, 62]). Therefore, simple and straightforward approaches for creating personalized models, which depend on large amounts of ground truth labels for training, cannot be applied. Recent advances in machine learning on training sparse models attempt to tackle this issue [35]. However, these methods lack interpretability. In addition, we show that this method does not perform well on our mobile sensing datasets.

Instead, we propose a method that effectively uses the study population's large volume of behavioral feature data to enable personalization. Our method is inspired by collaborative-filtering [30], initially used in recommendation systems, where a new item is recommended to a user based on other users whose preference are similar to this user. Following this intuition, we propose a *behavior relevance* metric that leverages individual users' behavioral similarities and differences to impute missing data and generate classifications. Figure 1 shows an overview of our data pipeline.

As a case study, we applied our method to the problem of depression detection using a mobile sensing dataset collected from first-year college students with low data-missing rates. To evaluate the performance of our method, we compare it against several baseline machine learning models. Our proposed method achieves an accuracy of 0.824 and an F1 score of 0.855. Compared to the state-of-the-art model [62], our model has higher performance by 5.1% and 5.5% respectively, with statistical significance ($p < 0.05$). The improvement still holds when we apply our method to a second dataset separately collected from a different college student population at a different institution (3.6% in accuracy and 2.8% in F1 score), thus demonstrating the pipeline-level generalizability of our approach (*i.e.*, applying the whole pipeline to another dataset). We also tested model-level generalizability (*i.e.*, using one dataset to train the model and the other dataset as the testing set), which had an accuracy of 60.8%, and needs further investigation in future work.

Beyond achieving better classification results than existing machine learning methods, our behavior relevance metric can further be combined with association rule mining to generate individual behavior rules that provide personalized interpretability. On the depression detection dataset, the behavior rules extracted by our method can capture behavior differences between the students with and without depression that are supported by the existing literature, provide an individualized understanding of the behaviors of students with depressive symptoms, and suggest potential directions of intervention designs for depressive symptoms improvement.

The contributions of our paper are as follows:

- We present a new approach to leverage behavior relevance (including both similarities and differences) among users for personalized classification on mobile sensing data.
- We apply our method on depression detection among college students. Compared to the state-of-the-art model, our method is better by 5.1% in accuracy and 5.5% in F1 score ($p < 0.05$).
- We verify our method by replicating the pipeline on a second independent dataset. The results show an average improvement of 3.4% across performance metrics.
- We demonstrate that our method can generate personalized behavior rules for students with depressive symptoms that are highly interpretable and informative for potential mental-health improvement.

2 BACKGROUND

We first review the existing work on behavior capturing and modeling from mobile sensing data. In particular, we focus on depression detection as this is the application domain we use throughout the paper (Section 2.1). Our method aims to address personalized behavior modeling, thus we also review the related work in personalized machine learning models (Section 2.2).

2.1 Capturing and Modeling Human Daily Routine Behavior and Depression Status

Daily routine behaviors, such as movement patterns, sleep patterns, social activities, and physical activities, can be tracked by sensors embedded in mobile phones and wearable devices. Researchers have demonstrated the feasibility of using mobile sensing to capture and model daily behavior in many settings [25, 33, 59]. Major depressive disorder (MDD), also known simply as depression, is a common but significant health challenge. Research has found that depression affects approximately 216 million people globally [57]. In 2018, an estimated 7.2% of all U.S. adults had at least one depression episode over the past year [43]. The number increases up to 13.8% among young adults, with 8.9% having severe impairments. Detecting depression at an early stage can help mitigate or prevent its negative consequences. There is a growing realization that everyday devices, continuously and passively collecting behavioral data, can help us to understand the relationship between people's daily behaviors and symptoms of depression [7, 12, 33]. Successes in the last decade of using mobile sensing for depression-related research have made this topic increasingly popular. Earlier work focused on understanding the statistical relationship between depressive symptoms and features extracted from mobile sensing data [10, 29, 47]. For instance, Saeb *et al.* [47] identified a significant correlation between depression scores and location features (location variance, location entropy, and circadian movement), and also phone usage features (usage duration and frequency). Ben-Zeev *et al.* [10] observed a significant correlation between changes in depression scores and sleep duration, speech duration, and mobility.

More recently, researchers have compiled the findings from correlation analysis to create machine learning models for depression detection. For example, Farhan *et al.* [20] used location features to detect biweekly depression and their best model achieved an F1 score of 0.82 on a dataset with 79 college students over eight months. Wahle *et al.* [58] trained models on multiple data streams, including location, physical activity, phone usage, and WiFi scans, and achieved an accuracy of 61.5% for depression detection on a dataset with 36 participants over ten weeks. Wang *et al.* [60] hand-crafted several cross-sensor features from mobile and wearable data. Their best model achieved 81.5% recall and 69.1% precision on a dataset collected from 68 college students over two nine-week terms. Association rule mining [3] (ARM) is a powerful method for extracting interpretable behavior rules. Xu *et al.* [62] proposed an automated cross-stream feature extraction pipeline that leveraged ARM behavior rules to extract features from mobile sensing data. Their model achieved an accuracy of 81.8% and an F1 score of 84.3% on a dataset containing 138 college students over 16 weeks.

Most of the work in depression detection follows a one-size-fits-all or *population modeling* approach, *i.e.*, training one classification model for all testing samples. This approach often results in reduced model performance compared to what might be possible with personalized models. Moreover, even if a population model did perform well and was interpretable (*e.g.*, [60, 62]), it only provides an understanding of the population behavior as a whole. However, no two people are identical. Each individual behaves differently from others. A good interpretable classification model should take this level of difference into account. Some previous work has trained individual or *personalized* models on an individual's own data. For example, Canzian and Musolesi [12] used one person's location features to train an individual-level depression detection model. Their model achieved 0.71 sensitivity and 0.87 specificity scores. However, isolating one individual from the rest of the population will ignore meaningful information that comes from others, *e.g.*, those who are similar to or very different from this individual. Our method, inspired by memory-based collaborative filtering, leverages the behavior correlations among individuals to predict whether students are experiencing symptoms of depression, and ARM to generate individualized behavior rules to provide detailed understanding. Our results show that such a method can achieve better prediction performance than [62] and more personalized interpretation.

2.2 Personalized Machine Learning and Behavior Modeling

There are two major types of personalized machine learning algorithms: sample-specific methods and similarity-based methods [56]. We review both classical and modern examples of the two methods. Sample-specific methods are model-based methods that leverage training and testing samples' features to enable personalization when inducing a model. Classical examples include lazy decision trees (LDT) [22] and lazy Bayesian rules (LBR) [66], where part of the training occurs when the testing sample is collected. The aforementioned case of building separate models on individual data [12] is another example, where the individual identifier is used to select the model (*i.e.*, the model trained on this individual's data) for prediction. There have been more advances in machine learning community [49, 63]. For example, Lengerich *et al.* [35] proposed personalized logistic regression model (PLR) to include different parameters for every sample, with the parameter matrix having a low-rank property as the constraint on the parameters' degree of freedom.

Table 1. Summary of approaches found in closely related work. *Method* describes the approach and gives examples from the literature. The other three columns indicate the properties of each method. For methods that can be trained on few labels per person, we choose typical ones as baselines (marked by *).

Method	Interpretability	Personalization	Number of Labels Per Individual
Current passive sensing methods for depression detection [20, 58, 60], **[62]***	Yes	No	Few
General sample-specific method *e.g.*, LDT [22], LBR **[66]***, PLR **[35]***	No	Yes	Few or Many
Sample-specific method for behavior modeling [12, 44, 65]	Limited	Yes	Many
General similarity-based method *e.g.*, LWR [6], KNN **[30]***	Limited	Yes	Few or Many
Similarity-based method for behavior modeling [1, 34, 36, 52]	Limited	Yes	Many
Our method	Yes	Yes	Few

In the area of behavior modeling, there is some related work on activity recognition and affective computing that has leveraged large-scale data, thus supporting sample-specific methods. For instance, Zhao *et al.* [65] proposed a semi-supervised activity recognition method. They first used a general decision-tree on the testing user's unlabelled data, and then added a one-step K-means clustering to re-assign the outputs so that an individual-specific tree can be retrained. Rudovic *et al.* [44] leveraged individual demographics and behavioral assessment scores to train a hierarchical deep learning model for emotion recognition. However, these sample-specific methods usually require a large number of ground-truth labels for each individual. In the area of passive sensing, the collection of labels is expensive, and many data sets have only one label for each individual (*e.g.*, [59]). Further, these methods do not focus on interpretability, which is an important factor for behavior understanding.

In contrast, similarity-based methods use a similarity or a distance measure and combine training samples in some fashion for the prediction. Traditional methods such as K-Nearest Neighbour (KNN) [30] and locally weighted regression (LWR) [6] are examples of this approach. There have been more recent advances in the behavior modeling area with this type of method. Lopez-Martinez *et al.* [36] used spectral clustering to group individuals into several groups based on their behavioral profiles and then applied multi-task learning for subjective pain estimation. Sun *et al.* [52] proposed a multi-task learning objective function for activity recognition by including the similarities of the activity patterns between activity pairs. Lane *et al.* [34] leveraged personal informatics, mobility behavior, and raw-sensor-data to construct three similarity matrices. They used the three matrices to initialize parameters when training three boost models for activity recognition. Then, they leveraged majority voting of the three models to determine the final predictions. Abdullah *et al.* [1] built on this idea. They added a clustering step based on behavioral similarity and only conduct the training within clusters for activity recognition. Their method assumed there is a large number of labels from each individual, however this is not always practical, especially in other behavior modeling areas such as depression detection. In addition, they did not include the temporal behavior information into the similarity measurement, which is important for behavior understanding.

To summarize, in the space of behavior modeling, there is ample evidence that behavior activity traces capture information that can be used for detecting important facets of the human experience. However, most methods follow a population-modeling approach, despite great variability in behavior between people and over time.

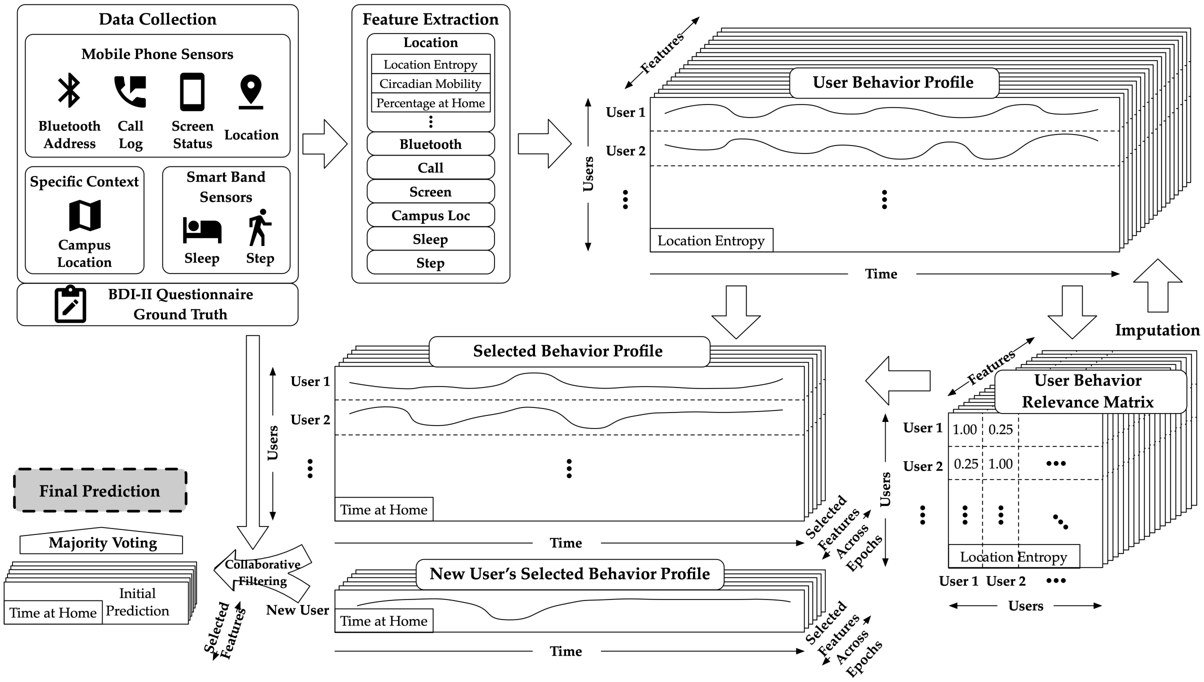


Fig. 1. The overview of the pipeline

In many behavior modeling areas such as depression detection, sparse labels are captured through the effort of participants, making a per-person model infeasible. Even in areas such as activity recognition where it is easier to obtain a rich set of ground truth labels, the existing personalization methods do not provide enough interpretability. Table 1 summarizes some closely related work. Our method addresses both personalized behavior detection (*e.g.*, depression detection) and personalized interpretation (*e.g.*, understanding the life experience of a person with depressive symptoms). We show that memory-based collaborative filtering can be used in the domain of behavior modeling. We enhance the interpretability of our method by leveraging ARM to generate human-understandable behavior rules. We describe this approach in more detail in the next section.

3 PERSONALIZATION ALGORITHM

We now present our method that leverages behavioral relevance for personalized classification (Section 3.1). We further introduce a behavior rule mining process to provide individualized interpretation (Section 3.2). The core idea is to leverage a behavior relevance metric to identify a unique, informative sub-group of other users in the dataset. The personalized group is then leveraged to obtain classification and interpretation results.

3.1 Classification

We first introduce the concept of user behavioral profiles and its relationship with collaborative-filtering (Section 3.1.1). Then, we leverage the correlation matrix from user behavioral profiles to impute missing data (Section 3.1.2). We then propose a measurement of the behavior relevance (square of the correlation) using the imputed behavioral profiles (Section 3.1.3) and use the metric to select features that have good performance in the training set (Section 3.1.4). When a new testing user is added to the analysis, we employ the selected features to generate intermediate classification outputs. Finally, we use majority voting to compile the intermediate outputs into the final classification output (Section 3.1.5). Figure 1 visualizes the overall pipeline.

3.1.1 Collaborative Filtering and Behavior Relevance Metric. Our method is inspired by the idea of memory-based collaborative filtering [30]. It was originally used in recommendation systems, where an item is recommended to a user based on a certain similarity between two users or items, *i.e.*, a item from another user who has similar preferences for items as this user (user-level), or another item that has a similar preference profile as this item (item-level). Borrowing the idea of the user-level collaborative filtering, we propose the concept of a *user-behavior profile* to depict a group of users' behaviors. Each user-behavior profile, represented by a matrix (see Figure 2), focuses on one particular feature. The user-behavior profile can be viewed as a user-item matrix from traditional collaborative filtering. Our method looks at each feature independently. Therefore, we also include users' target labels (*i.e.*, ground truth labels) in each behavior profile (the bold frame marked with L in Figure 2), which can be regarded as a column of "special items" in the profile matrix. A new user's "special item", marked by X, is the element that needs to be predicted. Note that the matrix will inevitably contain missing values (marked as ?) due to software, hardware and/or user issues that result in missing data.

3.1.2 Data Imputation. After constructing the user behavior profile for each feature, we then impute the missing data in the behavior profiles. Following Xu *et al.* [62], we first normalized each user's features by discretizing all behavior features into three levels – low as 1 (0-33 percentile), medium as 2 (33-66 percentile), and high as 3 (66-100 percentile) – using their own data as each individual's behavior has its own consistency and set of routines. This can also effectively reduce the bias of outliers. Then, we use the weighted average of other users' normalized data as the imputed value, where the weights are the correlation of the longitudinal feature value between this user and other users. The intuition is to leverage the data from people who have similar behavior (for the feature that the data represents) to impute the missing value. Algorithm 1 lists out the detailed steps for data imputation.

3.1.3 Behavior Relevance Metric. From the imputed data, we propose a *behavior relevance metric*. Prior work focuses on people who have similar behaviors (*e.g.*, [1, 34]). However, when predicting the final outcome (*e.g.*, having depressive symptoms or not), the scope can be expanded. For people whose behaviors are strongly

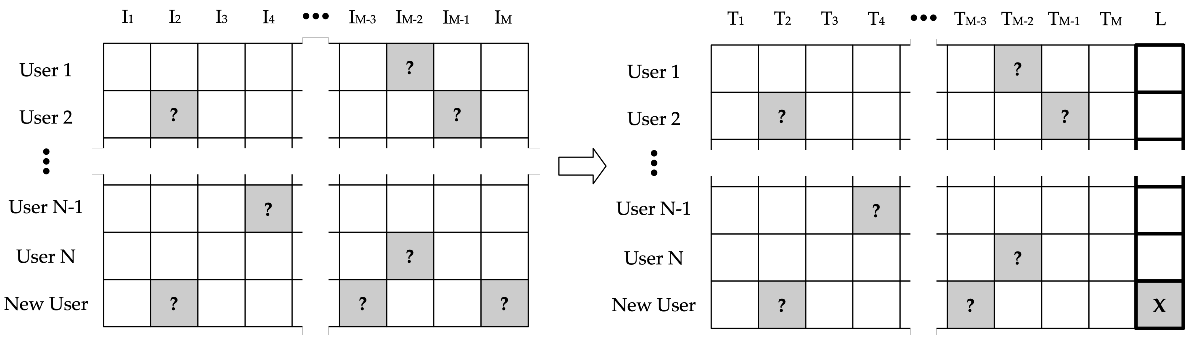


Fig. 2. The similarity between our method and collaborative filtering. The left part shows a **user-item** matrix that is commonly used in recommendation systems, where the "?" marks indicate preference scores to be predicted. In contrast, the right part shows the common data format in mobile sensing. The matrix is the **behavior profile** of one particular feature (each user has a time series data from T_1 to T_M), plus a column of labels (Column L). where "?" marks indicate the missing value and "X" mark indicate the target label to be predicted, *e.g.*, whether the new user is depressed or not.

related to a target user, they can be divided into to two types. People who behave similarly (with strong positive correlation) are the ones whose behavior patterns change in a similar way as the target user, *e.g.*, they all have more phone calls during the weekend. Moreover, there are also people who behave differently (with strong negative correlation). Here, behavior patterns change in a way opposite to the target user, *e.g.*, the target user mostly stays at home and moves less on weekday evenings, while these people often hang out for socializing at that time. Both types of people are relevant to a target user. Their similarities and differences may be indicators for classification (*e.g.*, social behaviors are related to depressive symptoms), thus both should have representation. Therefore, we further define the *behavior relevance metric* as the square of the Pearson correlation coefficient. The metric indicates high relevance if two users have either very similar or very different behavior.

3.1.4 Feature Selection. The number of behavior features captured by mobile phones and wearable devices can be huge. Moreover, prior literature has suggested that people have different behavior patterns during different times of the day [13], and between weekdays and weekends [46]. Similar to Wang *et al.* [59], we group raw sensor data into 10 epochs that capture behavior at different times: five epochs for weekdays (morning 6am-12pm, afternoon 12pm-6pm, evening 6pm-12am, night 12am-6am, and the whole day) and the same for weekends. This further increases the number of behavior features by an order of magnitude. Therefore, we need to select the features that are the most helpful for distinguishing target labels.

```
Data:
$E$: the epoch set; $D$: the days in the dataset; $U$: users in the training set;
$F$: the overall feature set. $F_e$: the feature set of a particular epoch $e$ ($\subseteq E$), $F_e \subseteq F$;
$R_{E,F}$: the list of raw behavior profiles. Each matrix $R_{e,f}$ ($|U| \times |D|$) is the behavior profile of feature $f$
  ($\subseteq F_e$) in epoch $e$ ($\subseteq E$). $R_{e,f}$'s rows and columns can be indexed by a number or a list. For example,
  $R_{e,f}[u,d]$ locates the feature value of user $u$ on day $d$. $R_{e,f}[u,D]$ locates the feature array of user $u$ in
  days $D$. The same can be applied to other matrices in algorithms;

$P_{E,F} = Copy(R_{E,F})$ ;                          // The placeholder of the imputed behavior profiles
for $f$ in $F$ do
    $e = GetEpoch(f,E)$;
    $Cor_U = PairwiseCor(R_{e,f}, R_{e,f})$ ;         // Calculate the user-pairwise correlation matrix
     (|U|x|U|) with the missing value ignored
    for $u$ in $U$ do
        $D_m = FindDayMissing(R_{e,f}[u,D])$ ;        // Get the days where u has missing data
        for $d$ in $D_m$ do
            $U_{nm} = FindUsersNotMissing(R_{e,f}[U,d])$ ;        // Find users who have data on d
            $weight_u = Cor_U[u, U_{nm}]$ ;                      // Use correlation scores as weights
            $P_{e,f}[u,d] = WeightedAvg(R_{e,f}[U_{nm},d], weight_u)$ ;      // Impute the missing data
        end
    end
end
$Return(P_{E,F})$ ;                                   // Return the imputed behavior profiles
```

Algorithm 1: Data Imputation.

```
Data:
$E, F, F_e, D, U$ same as Algorithm 1;
$L$: the label list in the training set. $|L| = |U|$. The list can be indexed by a user $u$ to get the label $L[u]$, or by
  a list of users $U$ to get the label list $L[U]$. The same can be applied to other lists/arrays;
$P_{E,F}$: the list of imputed behavior profiles. Each matrix $P_{e,f}$ ($|U| \times |D|$) is the imputed behavior profile of
  feature $f$ ($\subseteq F_e$) in epoch $e$ ($\in E$);

$RankingScore_F = EmptyArrayWithSize(F); Threshold_F = EmptyArrayWithSize(F)$ ;
for $u_{vd}$ in $U$ do
    $U_{tr} = U \setminus \{u_{vd}\}$ ;            // Evaluate across each training user to ensure stability
    $accs = EmptyArrayWithSize(F)$;
    for $f$ in $F$ do
        $e = GetEpoch(f, E)$;
        $Cor_{U_{tr}} = PairwiseCor(P_{e,f}[U_{tr}], P_{e,f}[U_{tr}])$ ;   // Calculate user-user correlation matrix
        $Rel_{U_{tr}} = Cor_{U_{tr}} \odot Cor_{U_{tr}}$ ;                // Correlation square as the relevance matrix
        $weights = Rel_{U_{tr}} - diag(Rel_{U_{tr}})$ ;                       // Relavance metrics against others
        $labelscores = WeightedAvg(L[U_{tr}], weights)$ ;      // Label scores calculated from others
        $th1 = Avg(labelscores[\{u; u \in U_{tr}, L[u] = T\}]); th2 = Avg(labelscores[\{u; u \in U_{tr}, L[u] = F\}])$;
        $th = Avg(th1, th2)$ ;                                                // Splitting threshold
        $UpdateAvg(Threshold_F[f], th)$;
        $accs[f] = AccuracyByThreshold(labelscores, th, L[U_{tr}])$;
    end
    $Filter(accs, th = 0.5)$ ;       // Remove features that perform poorly on the training set
    for $f$ in $F$ do
        if $Rank(f, accs) \in TopRank(accs)$ ;                               // Assign ranking scores
        then $RankingScore_F[f] += Rank(f, accs)$ ; else $RankingScore_F[f] += 0$ ;
    end
end
$SF = SelectTopFeatures(RankingScore_F)$;
$TH = Threshold_F[SF]$;
$Return(SF, TH)$ ;      // Return the selected features and their corresponding thresholds
```

Algorithm 2: Feature Selection

Using the behavior relevance metric, we conduct a feature selection process on the training set. For each feature, an inner leave-one-user-out loop is used to find the most important and stable features. Specifically, we take one user within the training set as the "validating user" each time (the rest are "training users"), and compute *label score*s for all training users, by calculating the weighted-average of the label value (False is -1 and True is 1), with the relevance scores (against other training users) as weights. Then, we calculate the average of these scores among users with False labels and another average among users with True labels. We use the mean of the two average scores as the splitting threshold. To see how well this feature works, its threshold is compared against each training user's label score to get a tentative label. Having the labels and the ground truth, we can

```
Data:
$E, D, U, L$ same as Algorithm 1;
$SF$: the selected feature set from Algorithm 2; $SF_e$: the selected feature set of a particular epoch $e$ ($\subseteq E$);
$TH$: the threshold lists corresponding to the selected features from Algorithm 2;
$P^U_{E,SF}$: the list of behavior profiles of training users $U$. Each matrix $P^U_{e,f}$ ($|U| \times |D|$) is the behavior profile
  of feature $f$ ($\subseteq SF_e$) in epoch $e$ ($\in E$);
$u_t$: a testing user; $P^{u_t}_{E,SF}$: the list of behavior profiles of the testing user $t$. Each matrix $P^t_{e,f}$ ($1 \times |D|$) is the
  behavior profile of feature $f$ ($\subseteq SF_e$) in epoch $e$ ($\in E$).

$Results = EmptyArrayWithSize(SF)$;
for $f$ in $SF$ do
    $e = GetEpoch(f, E)$;
    $Cor_{u_t} = PairwiseCor(P^{u_t}_{e,f}, P^U_{e,f})$ ; // Calculate the correlation matrix (1x|U|) between the testing users and users in the training set
    $Rel_{u_t} = Cor_{u_t} \odot Cor_{u_t}$ ;                  // Correlation square as the relevance scores
    $U_t = FilterUsers(Rel_{u_t})$ ;       // Remove users whose similary is among bottom-quartile
    $weight_t = Rel_{u_t}[u_t, U_t]$;
    $score = WeightedAvg(L[U_t], weight_t)$;
    if $score > TH[f]$ then $Results[f] = TRUE$ else $Results[f] = FALSE$
end
$FinalResult = MajorityVoting(Results)$ ;                   // Majority voting across epochs
$Return(FinalResult)$;
```

Algorithm 3: Memory-based Classification

obtain an average accuracy for this feature from training users. We filter out features whose validation accuracy is below 0.5 and assign a ranking score for the top ten percentile among the remaining features according to the accuracy value (score n for n^{th} best feature) and zero for other features. We repeat this across each "validating user" and get a series of ranking scores for each feature. We then sum the score and pick half features with the lowest scores (*i.e.*, top five percentile features) as the best features.

3.1.5 Majority Voting. When a testing user is added to the analysis (with already collected data), we first calculate their relevance scores (against the users in the training set) for only the selected features. Then, for each selected feature, we filter out the training users whose relevance score is among the bottom-quartile (*i.e.*, bottom 25 percentile, a conservative threshold [27]) as their behavior is not relevant to the new user and can introduce noise. This leads to a unique, personalized training set for each new user. For each selected feature, we calculate a weighted-average label score for the new user, and obtain an intermediate classification output using the splitting threshold calculated from users in the training set. In other words, for each selected feature, using data from the remaining users for that feature, we produce a classification result just for that feature. Finally, we use majority voting approach to aggregate these features' intermediate outputs, as shown in Algorithm 3.

3.2 Interpretation

In Section 3.1, we show how our behavior relevance metric can be used to select effective features to generate classification results. However, just having these effective features is not enough. We need to know more about the contexts to better understand an individual's behaviors. For example, the duration of phone usage may be an important feature for depression detection. But participants may interact with their phones at home versus at social places with different frequencies. Moreover, people have distinctive daily routines, *e.g.*, some may spend more time on their phones when they are at home while others may be more likely to use phones at social places. Such an interpretation needs to be personalized to obtain an accurate understanding that is tailored to each person. Beyond classification, we further propose a method that combine the relevance metric with association rule mining (ARM) to provide personalized interpretation.

Previous work applied ARM on the whole participant group to generate popular behavior rules among the population (*e.g.*, [62]). In contrast, we propose to focus on a single user's data for personalized interpretation. Our interpretation focuses on generating personalized behavior rules that can capture the behavior differences between target users and other users, provide more insights into their life experiences, and suggest potential directions on how to support behavior changes to achieve a desired goal.

As the interpretation is focused on behavior distinctions, we propose to identify a small subset of users whose behaviors are very different from a target user (Section 3.2.1). Then, we leverage ARM to mine frequent behavior rules separately, once on the target user's data and then on the identified users' data (Section 3.2.2). Finally, we identify the rules that can provide the most meaningful information (Section 3.2.3).

3.2.1 Identifying Informative User Groups with Negative Correlated Behavior. Different users have different degrees of similarities when compared to a target user. In order to generate personalized interpretation, we need to first identify a group of users that are the most informative. Users in the group have differing labels than the target user, and very different behavior on the selected features. For instance, if the target user is a student with depressive symptoms, then the identified group would be the subset of the users who do not have depressive symptoms and their behavior features are strongly relevant to the target user, but in a negative direction, *i.e.*, strong negative correlation (among the top-quartile for a given behavioral feature). It is worth noting that this process is conducted on each target user and each selected feature individually. Therefore, the identified groups are personalized to every user.

3.2.2 Behavior Rule Mining. We will use one target user t and one selected feature f_s, together with the identified user group g as the examples when introducing the next two steps. Given the target user t, we create an identified user group g in terms of the selected feature f_s. We then employ ARM on the discretized features two times to mine frequent behavior rules, once using the target user t's data and again using the identified group g's data. The output rules of ARM are in the form of $X \rightarrow Y$ with support $P(X)$ and confidence $P(Y|X)$, where both X and Y are a set of discretized features at certain levels.

Each selected feature f_s belongs to a particular epoch. The rule mining is performed on the whole feature set within the epoch, which outputs a large number of behavior rules. An informative rule should suggest meaningful behavior changes to influence the final outcome. To identify these rules, we only focus on the rules whose Y includes the selected feature f_s, because f_s affects the classification result during the majority voting procedure. We dynamically adjust ARM support and confidence thresholds to make sure the numbers of behavior rules from the target user and the identified group are no less than ten thousand.

3.2.3 Behavior Rule Pairing. Having the rules from the target user side t and the identified group side g, we need to make the two sides comparable. We propose a rule pairing approach to align the rules from the two sides.

If two rules (one from each side) have exactly the same antecedent and a similar consequent, they can be aligned. Specifically, two rules will be paired if they have identical X (the same features at the same discretized value levels) and the same features in Y (but not necessarily the same level). For example, if the target user has a rule R_t as $X_t : \{f_x(low)\} \rightarrow Y_t : \{f_s(medium)\}$, a rule to be paired on the identified group R_g needs to have X_g

```
Data:
$E$, $F$, $F_e$, $D$, $U$, $L$ same as Algorithm 1; $SF$: the selected feature set from Algorithm 2;
$u_t$: a target user with $L[u_t] = target$; $U_{ntar}$: non-target users with $L[U_{ntar}] \neq target$, e.g., with vs. without
  depressive symptoms;
$P_{E,F}$: the list of behavior profiles. Each matrix $P_{e,f}$ ($|U| \times |D|$) is the imputed behavior profile of feature $f$
  ($\subseteq F_e$) in epoch $e$ ($\in E$).

$PersonalizedRules = EmptyList()$;
for $f$ in $SF$ do
    $e = GetEpoch(f, E)$;
    // Get the identified group whose behavior are most negatively correlated
    $P_{u_t} = P_{e,f}[u_t, D]; P_{U_{ntar}} = P_{e,f}[U_{ntar}, D]$;
    $Cor_{u_t} = PairwiseCor(P_{u_t}, P_{U_{ntar}})$ ;    // Calculate correlation scores between each target
      user and non-target users
    $Rel_{u_t} = Cor_{u_t} \odot Cor_{u_t}$;
    $filter(Rel_{u_t}, Sign(Cor_{u_t}) < 0)$ ;                    // Focus on users with negative correlation
    $U_{idt} = GetTopUsers(Rel_{u_t})$ ;       // Get the top quartile users as the identified group
    // Mine behavior rules seperately, focusing on rules with f in Y
    $P'_{u_t} = \{P_{e,f}[u_t, D]; f \in F_e\}; P'_{U_{idt}} = \{P_{e,f}[U_{idt}, D]; f \in F_e\}$ ;           // Get full behavior set
    $BehaviorRules_{u_t} = AssociationRuleMining(P'_{u_t}, f)$;
    $BehaviorRules_{U_{idt}} = AssociationRuleMining(P'_{U_{idt}}, f)$;
    // Focus on the rules with the highest confidence under each context, i.e., X
    $BehaviorRules_{u_t} = UniqueContext(BehaviorRules_{u_t})$;
    $BehaviorRules_{U_{idt}} = UniqueContext(BehaviorRules_{U_{idt}})$;
    $PairedRules = Pair(BehaviorRules_{u_t}, BehaviorRules_{U_{idt}})$ ;                // Same context X
    // Select the top three rules with largest gap on rule confidence
    $TopRules = GetTopRules(PairedRules)$;
    $Append(PersonalizedRules, TopRules)$;
end
$Return(PersonalizedRules)$;
```

Algorithm 4: Personalized Interpretation

exactly same as X_t, *i.e.*, $X_g : \{f_x(low)\}$. Meanwhile, its Y_g needs to have the same features f_s, but not necessarily at the same level, *i.e.*, $Y_g : \{f_s[at\ any\ level]\}$.

On each side, it is possible that among the selected rules, there might be multiple rules having identical X and same features in Y, but at different feature value levels. Continuing the example, on the identified group side, one rule R_{g1} is $X_g : \{f_x(low)\} \longrightarrow Y_g : \{f_s(medium)\}$, while another rule R_{g2} is $X_g : \{f_x(low)\} \longrightarrow Y_g : \{f_s(high)\}$. This will lead to multiple pairs of rules, *i.e.*, R_t can be paired with both R_{g1} and R_{g2}. However, R_{g1} and R_{g2} appear with different frequency in the dataset, as represented by their confidence values (their support values are the same because of the same X_g). Including both pairs will introduce additional noise. Therefore, for each X, we only retain the rule with the highest confidence and discard other rules, as this rule is the most representative and indicates the most common behavior when X appears. Then, we pair these representative rules following our description above and discard the unpaired rules. Once we pair the rules, we select the top three rule pairs that have the biggest confidence gap between the two sides for interpretation. This step finds the largest behavior differences between the target user and the identified group. We repeat the process for each selected features to obtain a set of personalized behavior rules for the target user.

The pipeline is described in Algorithm 4. Note again that the whole pipeline is conducted on each target user independently. Thus the interpretation provided by the final selected rules is personalized.

4 DATA COLLECTION AND IMPLEMENTATION

We describe the two depression datasets that we use to demonstrate the effectiveness and the generalizability of our method (Section 4.1). We then briefly explain the feature extraction procedure (Section 4.2). In addition, we introduce how we implement our method with these datasets (Section 4.3).

4.1 Data Collection

Our data collection studies were inspired by and modeled after the work of Wang *et al.* [59]. The two datasets were collected from two separate Carnegie-classified R-1 universities in the United States in a very similar procedure. In both institutions, we got IRB-approval and recruited first-year undergraduate students *via* emails and Facebook posts. We ended up with 188 students in the first dataset (male 77, female 110, non-binary 1, age mean = 18.2, sd = 0.40), and 209 students in the second dataset (male 75, female 134, age mean = 18.4, sd = 0.69). In both studies, students were invited to a research lab to sign a consent form, download a mobile application to track sensor data from their smartphones, and get a Fitbit wearable device to track their steps and sleep behavior. They were asked to maximize the time that the app was alive in their phones' background, and wear the Fitbit over the whole study period. In the first institution, the study lasted one semester (spring semester, 16 weeks). At the beginning and the end of the semester, participants were asked to answer a well-established questionnaire, the Beck Depression Inventory-II (BDI-II) [9], to evaluate their depressive symptoms. The second institution follows a quarter system, and the study lasted two quarters (winter and spring), including the 1-week break in between quarters. In our study we used only the spring quarter data (10 weeks) as the second dataset. Participants were asked to answer the depressive symptom evaluation questionnaire at the end of the second quarter. Participants in both institutions were allowed to keep the Fitbit after the study and received up to \$205 and \$245, in each study respectively, for compensation, based on their compliance.

Table 2. Information of the two studies after removing students who dropped out or were missing a significant amount of data. Students with a last BDI-II score > 13 labelled as having depression, in accord with the interpretation of the BDI-II [9].

Study	Days	Overall Size	Dropped out or Removed	Dataset Size	iOS Users	Last BDI-II Outcome	
						Non-depression	Depression
Inst 1	106	188	50	138	97	59	38 (39.2%)
Inst 2	166	207	38	169	134	84	56 (40.0%)

4.1.1 Ground truth Collection.

In both institutions, we employed the BDI-II [9], a widely used psychometric test, for depressive symptoms severity measurement, to obtain ground truth. The BDI-II is one of the most widely used self-report measures of the presence and severity of depressive symptoms in non-clinical samples and clinical trials of depression [24]. Many studies have provided validation information and normative data about the BDI-II in college students in large sample sizes (*e.g.*, [16, 51, 61]). The outcome of the questionnaire ranges from 0 to 63. For college students, the cut-offs are 0-13 (no or minimal depression), 14-19 (mild depression), 20-28 (moderate depression), and 29-63 (severe depression) [19]. In both datasets, we labeled the students whose last (*i.e.*, end-of-term) BDI-II score was higher or equal to 14 as having depressive symptoms.

4.1.2 Passive Sensor Data Collection. Our data collection application was developed based on the AWARE Framework [21]. The application recorded a phone's nearby Bluetooth addresses, call logs, phone usage (charging activity and screen status), and location. Further, participants were required to wear a Fitbit that recorded their steps and sleep status (as shown on the left of Figure 1). Data from AWARE was anonymized locally on the phone and automatically transferred over Wi-Fi to our back-end server when the phone was being charged. Fitbit data was downloaded using the Fitbit Web API at the end of the study. Participants were asked to keep their phone and Fitbit charged and carry/wear them at all times during the study period. Table 3 summarizes the passive sensor data and their related behaviors.

After data collection, we removed users who did not have data for more than half of the study days, to avoid the bias of low-quality data. Although the BDI-II scores are not significantly different between the removed students and those of the retained students ($t = 1.84, p = 0.07$), such a removal could potentially introduce bias into our dataset. We will have more discussion about this in Section 6.4.

Table 2 summarizes the high-level details for the two studies. Previous work found that the data collected from the iOS platform and Android platform were quite different [37]. Our method focuses on the heterogeneity among individuals rather than the heterogeneity caused by hardware or software. Therefore, we focus on only the iOS platform users as they are the majority in both of our datasets. Among all iOS users, the rate of having depression is high in both institutions (39.2%/40.0%), which is similar to national rates for depression among college students ACHA-NCHA II [2].

4.2 Behavior Feature Extraction

As introduced in Section 3.1.4, we used an approach similar to Wang *et al.* [59] to group raw sensor data into epochs that capture behavior at different times of day. We grouped the data in five epochs for weekdays (morning 6am-12pm, afternoon 12pm-6pm, evening 6pm-12am, night 12am-6am, and all day) and the same five epochs on weekends, resulting in 10 epochs in total. We then aggregated sensor streams on a per-day, per-epoch basis into daily-epoch features, *e.g.*, the number of phone calls on the morning of Wednesday, February 14, 2018.

Most features were aggregated using a mix of mean, maximum, minimum, and standard deviation for sampled data, and count and duration for event-based data. However, some features required additional pre-processing.

Table 3. Sensor data and information aggregated into features.

Behavior	Feature Type	Source	Sampling	Information Being Aggregated into Features
Phone Usage	Screen	AWARE	event-based	Number of unlocks per minute, total time with interaction, total time unlocked
Social Activity	Call			Number and duration of in-coming /out-going/missed calls
	Bluetooth		1 per 10 minutes	Number of unique devices, number of scans of most/least frequent device
Mobility	Location			GPS latitude, longitude, altitude
	Campus			Location data integrated with the campus map, *e.g.*, classrooms, sport space, green space
Sleep	Sleep	Fitbit	1 per minute	Asleep/restless/awake/unknown duration and onset
Physical Activity	Step		1 per 5 minutes	Number of steps

For location features, we calculated location variance, total distance traveled, average/variance of speed, circadian movement [47], number of significant places visited, number of transitions between places, radius of gyration [12], percentage of time spent at top-3 frequent clusters, length of stay at clusters, and location entropy. We further looked into the relationship between the user's location patterns and the college campus map to capture more specific contextual information, focusing specifically on Greek houses (which tend to hold social events), residential halls, sports spaces, green spaces, and academic buildings. For Bluetooth, we clustered devices into frequently seen groups and count the prevalence of each cluster. For sleep, we added sleep efficiency, and sleep onset. As for steps, we identified active bouts and sedentary bouts according to their stepping behavior (less than 10 steps in a 5-minute interval).

Each sensor stream results in features that can capture behavior variability that might be influenced by depressive symptoms [5]. For instance, depression can cause sleep disturbances, which might impact features such as sleep duration [55], and diminished activities, which could be reflected in the number of steps taken [17].

4.3 Implementation of Our Algorithm

We first normalized each user's features relative to their own data. We then employed the normalized features to calculate the behavior relevance among users. After applying the imputation in Algorithm 1, we use the imputed data for feature selection and classification, following Algorithm 2 and Algorithm 3.

We applied the interpretation Algorithm 4 on each target user, *i.e.*, every student who was labeled as having depressive symptoms. Following the approach in Xu *et al.* [62], we first discretized each user's features relative to their own data (similar to the normalization step) into three levels: low, medium, and high. All of the features in the selected feature's epoch were input to ARM (as described in Section 4.2). The output rules of ARM were in the form of $X \rightarrow Y$. For example, a rule X : {*[Location] Home Stay Duration (high), [Activity] Step Counts (low)*} $\rightarrow Y$: {*[Screen] Phone Interaction Time (high)*} reflects that if a student were at home for a long duration and they did not walk around much, then they were likely focusing on their smartphone. This rule could reflect a common behavior on weekday evenings. As introduced in Section 3.2, we select rules based on each selected feature independently, during which we only focused on the rules whose Y contained this particular feature. Continuing the example above, if the selected feature were *[Screen] Phone Interaction Time*, then the rule would be retained, otherwise it would be discarded. In addition, among the retained rules, if there were rules that have identical X and only differ in levels of the features in Y, the rule with the highest confidence value would be retained. For example, one rule was $X_0 \rightarrow Y_1$: {*[Screen] Phone Interaction Time (high)*}, with confidence at 0.7, while the other rule was $X_0 \rightarrow Y_2$: {*[Screen] Phone Interaction Time (low)*} with confidence at 0.3. The rule $X_0 \rightarrow Y_1$ would be retained for pairing and $X_0 \rightarrow Y_2$ would be discarded. Following Algorithm 4, the final personalized rules were aggregated from all selected features.

5 EVALUATION

We evaluated the performance of our algorithm in a number of ways. Using the dataset from institution 1, we first compared our method against several traditional baselines and recent advances in Section 5.1. We further evaluated how much data our method needs in Section 5.2. Then, we verified the generalizability of our method using the second dataset in Section 5.3. Finally, we demonstrated how our method can generate personalized interpretation and inspected examples of personalized behavior rules in Section 5.4.

5.1 Personalization Leads to Higher Performing Models

We compared our method with a few closely related baseline methods on the same dataset (see Table 1). Some are typical/recent personalized machine learning methods. Some are popular time-series behavior modeling methods.

We implemented these baselines and throughout the evaluation, we employed leave-one-user-out cross-validation to avoid over-fitting [54, 64].

(1) Majority, a naive baseline that simply classifies all samples as the major class in the dataset.

(2) Single Best Threshold, a simple threshold-based method that used the best single aggregated feature (the mean and the variance over the whole study period of each epoch) as the splitting threshold.

(3) K-Nearest Neighbour (KNN), a typical similarity-based method that use similar neighbours for classification [30]. We adopted Euclidean distance over all features (the same as our model's input) as the similarity measurement. K was set as 5.

(4) Lazy Bayesian Rules (LBR), a classical sample-specific personalized learning algorithm that builds a naive Bayesian classifier specifically for each test sample when it appears [66]. We used all epochs' aggregated features for training, as this is the common practice for behavior modeling (*e.g.*, [62]).

(5) Long short-term memory (LSTM), a neural network that is commonly used to model intrinsic relationships in time series data [48], which shares similarity with our relevance metrics. Given the limited data size, we used a small two-layer bidirectional LSTM, both with 16 hidden units and each users' feature across the whole study period as one data point.

(6) Personalized Logistic Regression (PLR), a recent state-of-the-art sample-specific personalized algorithm [35]. It assigns specific parameters for each sample, with low-rank representation and external covariates as the approaches to limit the parameters' degree-of-freedom. We followed the practice in [35] to use the 2D t-SNE embedding of the aggregated features from the first half of the study as the external covariates, and the aggregated features from the second half as the training and testing features. We set the rank as 10 and the number of neighbors as 3 based on grid search.

(7) Multi-sensor Classifier (MSC), a popular method for depression detection that concatenates multiple sensors' average feature value and trains a classifier with off-the-shelf models. It closely replicates some previous work (*e.g.*, [20, 47, 58]). We used random forest since it is one of the most commonly used off-the-shelf models for passive sensing because of its robustness and good performance. The maximum depth and the tree numbers were set as 5 and 30 based on tuning.

(8) Contextually-Filtered Classifier (CFC), a state-of-the-art behavior modeling algorithm that identifies co-occurrence patterns among features and extracts contextually filtered features for model training [62]. Note that CFC is a population modeling approach. We followed the practices in [62] to extract contextually-filtered features and trained the model. The hyperparameters of the final AdaBoost decision-tree-based classifier, *i.e.*, maximum depth and the number of the estimator, were set as 5 and 20 based on tuning.

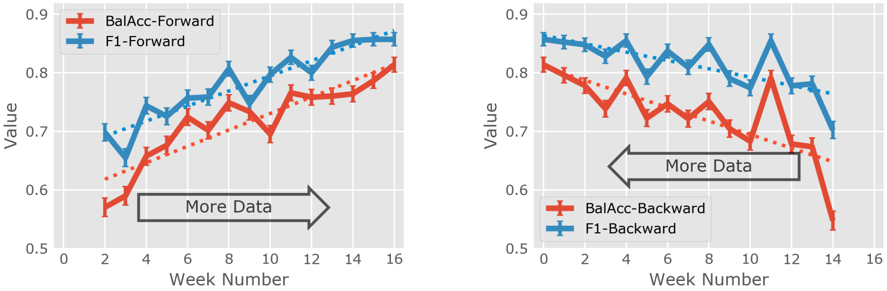
The results are summarized in Table 4 in terms of five metrics: accuracy (the overall success rate), balanced accuracy (taking the imbalance on labels into account), precision, recall and F1 score. Compared to the best-performing baseline model (CFC), our method's results have improvement in most metrics, particularly the accuracy (5.1%), the recall (10.1%) and the F1 score (5.5%).

5.2 The Amount of Data Needed by The Algorithm

We also investigated how much data is needed in order to obtain a satisfactory performance of our algorithm. We focused on several aspects that have important practical implications: How many days of data does the algorithm need to establish a good relevance metric (Section 5.2.1)? How many people are needed in the training set to build a high-performance model (Section 5.2.2)? Which feature types are important for the algorithm to work effectively (Section 5.2.3)?

Table 4. Comparison of baseline and state-of-the-art machine learning classifiers and our new algorithm. T-tests on both the balanced accuracy and the F1 score between our method and the best baseline CFC show that our method significantly outperforms the baseline method ($p < 0.05$ in all cases).

Classification	Accuracy	Bal Accuracy	Precision	Recall	F1 Score
Majority	0.608	0.500	0.608	1.000	0.756
Single	0.639	0.679	0.853	0.492	0.624
KNN [30]	0.567	0.527	0.627	0.712	0.667
LBR [66]	0.629	0.615	0.702	0.678	0.690
LSTM [48]	0.557	0.462	0.589	0.898	0.711
PLR [35]	0.667	0.668	0.765	0.661	0.709
MSC [20, 47, 58]	0.716	0.700	0.725	0.771	0.747
CFC [62]	0.773	0.781	0.863	0.746	0.800
Our Algorithm	**0.825**	**0.819**	**0.862**	**0.847**	**0.855**



(a) Counting Forward

(b) Counting Backward

Fig. 3. The results when using data from different numbers of weeks. Forward means using the data between the beginning of the data collection period (the beginning of the semester) and the particular week number, while backward means using the data between the particular week number till the end of the period (the end of the semester). Error bars indicate the standard error of the mean.

5.2.1 How Many Days Does The Algorithm Need? We evaluated the effect of the number of days of data from two perspectives: forward and backward. Given a particular week number, forward means only using the data between the beginning of a study and the week number, while backward means only using the data between the week number through until the end of the period, *i.e.*, the time when students finished the final BDI-II survey. These two perspectives are complementary. The first perspective indicates how early we can use the collected data to predict the depression status at the end of the semester, while the second perspective indicates how depression can be reflected from the most recent behavior. Figure 3 visualizes the results of both perspectives.

In general, both figures present an increasing trend on the two metrics as the number of days used for training increases (reading left to right in Figure 3a and right to left in Figure 3b). We observed some interesting phenomena. In the counting forward approach, it follows an overall trend that the more data we have, the better performance we can achieve. The performance of only using the data from the first several weeks to predict the depressive status at the end of the semester is not satisfactory. Moreover, we see a small drop in balanced accuracy from week 8 to week 10, accompanied by a smaller drop in F1 score (see Figure 3a). This was during the mid-term period and students might have been busy preparing for exams, indicating that such an break in their routine might affect the effectiveness of the relevance metric for depression detection.

From the backward perspective, there is a peak in balanced accuracy using five weeks of data (end of study back to week 11) (see Figure 3b). The F1 score also has a small peak, but it is less significant. This suggests that a one-month period could be a good time window for signalling depression status. We will have more discussion about the implications in Section 6.2.

5.2.2 How Many People Does The Algorithm Need? We also evaluated the method in terms of the number of users required to establish a good training set. To determine this, we uniformly sample a certain number of users from the whole dataset and call this the training dataset. The remaining users comprise the testing dataset. The process is repeated one hundred times to obtain the mean and the standard error. Figure 4 visualizes the results.

Not surprisingly, both the balanced accuracy and the F1 score increase monotonically as the number of users increases. The more users in the training set, the more likely a testing user can find users with similar or opposite behavior, leading to better results. Such an increase becomes slower when the number of users is above 60. Both metrics are close to a plateau when there are 60 participants.

5.2.3 How Does Each Feature Type Affect Algorithm's Performance? In order to investigate the effect of each feature type in Table 3, we further conducted a feature ablation study. For the seven features types – phone screen, call, Bluetooth, location, campus, sleep, and step (as shown in Table 3) – we removed one of them and re-ran the whole pipeline using the remaining six feature types. Figure 5 summarized the results.

We found that the two mobility-related feature types (location and campus) were the most important, and removing them leads to the biggest drop in the balanced accuracy (17.3% and 6.6% absolute value, respectively). This is supported by the previous literature [20, 47]. In contrast, removing Bluetooth feature has the least effect, with 1.4% absolute value drop.

5.3 Generalizability of The Algorithm

We evaluated the generalizability of our algorithm from two aspects: 1) the pipeline-level generalizability (Section 5.3.1): applying the whole algorithm to another dataset; and 2) the model-level generalizability (Section 5.3.2): using one dataset to train the model and another dataset from a different population as the testing set.

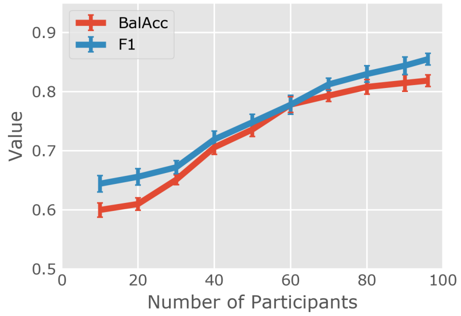


Fig. 4. The results when using different numbers of users for training. Each data point is the mean of one hundred random samples from the dataset. Error bars indicate the sample standard error.

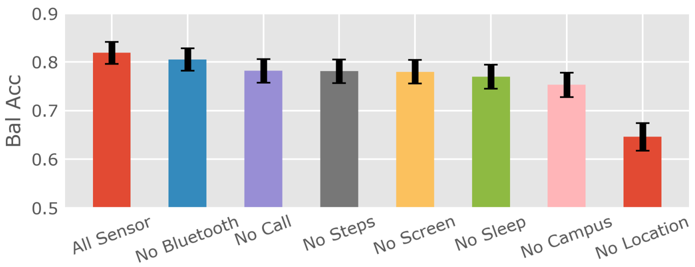


Fig. 5. The balanced accuracy results of the feature ablation study. Each time one of the seven feature types (as shown in Table 3) is removed and the whole pipeline is applied on the remaining features. Error bars indicate the sample standard error.

5.3.1 Generalizability of The Pipeline. We replicated the whole pipeline on the second dataset (see Section 4.1). We also compare the same baselines introduced in Section 5.1 against our method. All of the results are summarized in Table 5 with the same five metrics.

On the second dataset, our algorithm still outperforms other baselines. The method achieves an accuracy at 0.791, a balanced accuracy at 0.773, and a F1 score at 0.814, which has an advantage of 3.6%, 3.8%, and 2.8% compared to the best CFC baseline, respectively. This verifies the generalizability of our pipeline: our method has the potential to be applied to other independent studies.

Table 5. Evaluation of the pipeline-level generalizability of our method. We replicate the pipeline on another dataset with similar data format. T-tests on the balanced accuracy and the F1 score shows that our method significantly outperforms the best baseline method ($p < 0.05$ in all cases).

Classification	Accuracy	Bal Accuracy	Precision	Recall	F1 Score
Majority	0.600	0.500	0.600	1.000	0.750
Single	0.626	0.637	0.742	0.583	0.653
KNN	0.583	0.576	0.671	0.607	0.638
LBR	0.647	0.620	0.692	0.750	0.720
LSTM	0.536	0.452	0.575	0.869	0.692
PLR	0.710	0.682	0.731	0.819	0.773
MSC	0.732	0.719	0.759	0.821	0.789
CFC	0.755	0.735	0.778	0.833	0.805
Our Algorithm	**0.791**	**0.773**	**0.814**	**0.854**	**0.833**

5.3.2 Generalizability of The Model. Beyond the pipeline-level generalizability, a stronger level of generalizability would be model-level generalizability, *i.e.*, applying a trained model directly on a new dataset. We evaluate this aspect by using one dataset for training and the other one for testing.

Our relevance metric requires the two datasets to have the same number of days. Thus we tried a few methods for alignment: uniformly down-sampling the first dataset as a semester (16 weeks) is longer than a quarter (10 weeks); taking the last 10 weeks from the first dataset; and taking the last month from both datasets, according to the good performance in Figure 3b (the last month could get rid of the effect of midterms in both institutions). For each method, we evaluated the model-level generalizability from two directions: train on the first institution's dataset and test on the second, and vice versa.

However, among all alignment methods and directions, the best model - using the last month's data from institution 2 for training and testing on the last month's data from institution 1 - still has an unsatisfactory performance. It achieves an accuracy of 0.608, a balanced accuracy of 0.570, and an F1 score of 0.698. This poorer performance could be caused by multiple factors. For example, students from the two institutions are quite different. It might be hard to find someone from one institution whose behavior is relevant to a student from another institution with a different academic calendar system. Such a difference might become larger when considering environment differences into account, such as the institutional culture, city size, season of the year, *etc.* These factors would increase the difficulty of the model-level generalizability.

5.4 Interpretable Personalized Behavior Rules

Section 5.1 to Section 5.3 evaluates the performance of our classification algorithm described in Section 3.1. Beyond achieving good classification results, we further show that our method (Algorithm 4) is able to generate personalized understanding of individual students, especially those with depressive symptoms.

Table 6. Evaluation of the model-level generalizability of our method. A model is trained on one dataset and then tested on another dataset collected from another population.

Method	Train/Test	Accuracy	Bal Accuracy	Precision	Recall	F1 Score
Uniform Down-sample	Ins1/Ins2	0.470	0.472	0.585	0.463	0.517
Uniform Down-sample	Ins2/Ins1	0.515	0.471	0.588	0.678	0.630
Last 10 weeks	Ins1/Ins2	0.567	0.569	0.676	0.561	0.613
Last 10 weeks	Ins2/Ins1	0.557	0.504	0.611	0.745	0.672
Last one month	Ins1/Ins2	0.597	0.547	0.642	0.768	0.700
Last one month	Ins2/Ins1	0.608	0.570	0.657	0.745	0.698

The behavior rules generated from our method are tailored for every student who have positive labels (*i.e.*, having depressive symptoms). Using the dataset from institution 1, we randomly picked two students with moderate depression (P18 with post BDI-II score 24, and P72 with post BDI-II score 26) as anecdotal examples and investigate how the personalized rules capture behavior differences from their identified user groups (students without depressive symptoms) with different behaviors. Table 7 lists out a subset of these rules.

The majority of the rules from the two students are different. It is expected as these two students did not have identical behavior. Interestingly, close inspections of the rules reveal both the homogeneity and the heterogeneity at the same time. Some of their behaviors share commonalities that are supported by existing depression-related literature, while some other behaviors are quite different between the two.

The top four rows of Table 7 shows examples of the homogeneity on sleep pattern and phone usage behavior. P18's weekend morning rule No.4 indicates that P18 had more interrupted sleep (higher number of sleep bouts) compared to their identified group on weekend mornings when they were not at social spaces or dorm, and their total numbers of sleep bouts (including being asleep, restless, awake) were low. Similarly, P72's weekday all day rule No.3 indicates that P72 had shorter sleep duration than the identified group when they have a low number of incoming calls and away from social spaces. Both rules indicate that the two students' sleep patterns were disturbed. Similar sleep patterns were consistent among other students. Among the 38 students who were labeled as being depressed, 92.1% of them (35 out of 38) had more than half of the rules about sleep showing a more disturbed sleep pattern than in their respective identified group. Moreover, 73.7% (28 out of 38) had more than seventy percent of the rules showing such a trend.

Beyond sleep patterns, homogeneity was also observed on phone usage patterns. For P18's weekday evening rule No.10, both the user and the identified group had high number of unlock per minute (same Y) when they mostly stayed at somewhere far from dorms and social spaces. Their rules have a similar support value but P18's rule had a higher confidence value, which means that this rule was more common for P18 than for the identified group. This suggests that P18 was unlocking their phone more often. Likewise, P72's weekday evening rule No.2 shows that P72 spent longer time than the identified group interacting with their phones when they remained sedentary at some place out of the dorm. Both rules reflect that the two students had more active phone usage than their respective identified groups. These patterns were also observed in other students: 73.7% of the users with depressive symptoms have the majority of the rules about phone usage showing the same trend. We have more discussion about the relationship between these behaviors and existing literature in Section 6.

In contrast, some rules capture behavior heterogeneity between P18 and P72. Examples related to mobility and communication behavior are shown in the last four rows in Table 7. P18's weekend all day rule No.23 suggests that P18 spent a shorter time in sports spaces (for exercise) and green spaces (for relaxation) than the identified group when they were out of dorms and far from large groups of people (indicated by the Bluetooth feature). However, P72 had a rule with the opposite behavior: P72 spent more time in sports spaces than the identified group when they were out of the dorm and had a large location variance. A similar contrast was also observed in phone call behavior. P72 had a shorter duration of outgoing phone calls than that of the identified group (weekend all day rule No.21), indicating less social communication, while P18 had more outgoing calls than the identified group under similar contexts. These distinguishing rules reflect individual behavior differences between P18 and P72.

As our interpretation method is designed to mine each individual's behavior rules independently, it can capture the behavior similarity and the difference among users at the same time. Examples in Table 7 support that our method can generate personalized rules that can support personalized interpretation.

Table 7. Examples of top paired rules that capture behavior differences between a target user with depression and their identified groups without depression. Two rules in a pair have identical X and the same selected feature in Y (shown in bold). Each item is displayed in a "[feature type]feature(discretized value)" manner. The bold feature highlights the difference between the target user and the identified group. The dashed lines group the type of selected features such as sleep-related and screen-related behavior. The first four rows show homogeneity between the two students, *i.e.*, their differences against the opposite groups are in the same direction, while the last four indicate heterogeneity. *e.g.*, P18 has more social communication than the opposite group but P72 has less.

PID	Rule	X	Y_{tar}	sup_{tar} $conf_{tar}$	Y_{oppo}	sup_{oppo} $conf_{oppo}$	Property
18	Wkend Morning No.4	- [Campus] Pct. of time at social space or dorm (low) - [Sleep] Total bout nums during the sleep (low)	- [Sleep] **Num of bouts being asleep** (medium)	0.200 0.667	- [Sleep] **Num of bouts being asleep** (low)	0.152 0.298	More disturbed sleep pattern
72	Wkdy Allday No.3	- [Campus] Pct. of time at dorm (low)	- [Sleep] **Duration of being asleep** (low) - [Campus] Pct. of time at sports space (low)	0.367 0.550	- [Sleep] **Duration of being asleep** (medium) - [Campus] Pct. of time at sports space (low)	0.172 0.246	More disturbed sleep pattern
18	Wkdy Evening No.10	- [Location] Moving time Pct. (low) - [Campus] Pct. of time at social space or dorm (low)	- [Screen] **Num of unlock per minute** (high)	0.132 0.588	Same as Y_{tar} with lower *conf*	0.137 0.413	More phone interaction
72	Wkdy Evening No.2	- [Location] Pct. of time at home (low) - [Step] Avg duration of sedentary bouts (high)	- [Campus] Pct. of time at dorm space (low) - [Screen] **Avg duration of interaction bouts** (high)	0.171 0.302	- [Campus] Pct. of time at dorm space (low) - [Screen] **Avg duration of interaction bouts** (low)	0.161 0.395	More phone interaction
18	Wkend Allday No.23	- [Bluetooth] Num of unique others' device (low) - [Campus] Pct. of time at social space or dorm (low)	- [Campus] Pct. of time at greens space (low) - [Campus] **Pct. of time at sport space** (low)	0.267 1.000	Same as Y_{tar} with lower *conf*	0.222 0.435	More time at sport space
72	Wkend Allday No.18	- [Location] Log of location variance (high) - [Campus] Pct. of time at dorm (low)	- [Campus] **Pct. of time at sport space** (low)	0.267 0.533	Same as Y_{tar} with higher *sup* and *conf*	0.356 0.744	Less time at sport space
18	Wkend Allday No.21	- [Call] Num of outgoing calls (low)	- [Sleep] Total bout nums during the sleep (low) - [Call] **Num of outgoing calls (medium)**	0.133 0.364	- [Sleep] Total bout nums during the sleep (low) - [Call] **Num of outgoing calls (low)**	0.300 0.720	More call communication
72	Wkend Evening No.7	- [Call] Num of outgoing calls (low) - [Campus] Pct. of time at dorm (low)	- [Call] **Duration of outgoing calls (low)**	0.166 0.625	Same as Y_{tar} with lower *sup* and *conf*	0.158 0.441	Less call communication

6 DISCUSSION

We first discuss the insights obtained from the personalized behavior rules and their relationship with the depression literature in Section 6.1. We then discuss the practical implications for depression detection leveraging the results of our analysis Section 6.2. We also discuss potential usage of our methods beyond depression detection in Section 6.3. Finally, we reflect on the limitations and potential future work in Section 6.4.

6.1 Personalized Rules and Depression Literature

Many of our findings in Section 5.4 are consistent with the existing literature on depression, adding support for the validity of our methods. For instance, the first two rules in Table 7 are about sleep behavior for students with depressive symptoms. P18's weekend morning rule No.4 reflects more sleep bouts during the sleep, and P72's weekday allday rule No.3 indicates shorter sleep duration. Although the rules in Table 7 are just example rules from two anecdotal cases, we found consistency among populations: 92.1% of the students with depressive symptoms have the majority of the rules reflecting less favorable sleep patterns. These results can be supported by relevant findings in psychology and clinical psychiatry that sleep disturbance, insomnia, and hypersomnia, are common symptoms of depression [5, 53, 55]. In addition, the third and fourth rules in Table 7 indicate the effects of depression on phone interaction behavior. P18's weekday evening rule No.10 indicates more frequent screen unlock behavior, and P72's weekday evening No.2 indicates longer interaction duration. Similar behaviors are also observed in other students having depression: 73.7% have the majority of rules showing more frequent phone usage patterns. These align with the rich literature that depression may lead to more phone usage [17, 23, 47, 58].

Rules showing homogeneity have been found in a good population model [62]. The value of our approach is that it further finds rules at a more fine-grained individual level: it finds rules that highlight individual differences, as demonstrated by rules showing heterogeneity among students. Four rows in the bottom of Table 7 shows examples for mobility and communication patterns. Although both P18 and P72 had depression, some of their behaviors still differed significantly. This also supports the intuition behind our method that training samples are treated differently according to their behavior relevance against the target user. Moreover, the interpretation rules suggest a new opportunity to leverage these personalized, contextual rules in technology-supported interventions for people with depression. The rules point out potential personalized design considerations for each individual so that an intervention can better match a particular user's behaviors. For instance, simply suggesting more movement and socializing should not be a universal intervention recommendation for all people with depressive symptoms. Some users should follow this while it is not appropriate for others. It would be unusual for a clinician to just rely on universal recommendations [42]. The personalized rules may help to guide a more nuanced assessment of individuals' needs and the intervention can be tuned to better fit each individual. Continuing the example of P18 and P72 in Table 7, the last two rules show the heterogeneity on phone call communication, indicating that it might be more effective to offer socialization suggestions to P72 than P18.

6.2 Amount of Data for Depression Detection

Figure 3b suggests alternatives other than using data from the whole study period to build a model. In particular, when using the data from week 11 to week 16 (the last month), the model achieves 0.790 in balanced accuracy and 0.851 in F1 score, which is quite close to the best performance when using the whole dataset (0.819 balanced accuracy and 0.855 F1 score,see in Table 4). Moreover, the last-month period also shows the strongest model-level generalizability across the two institutions (Table 6). These results suggest that for people with depressive symptoms, their recent-month behaviors may have more in common than the behaviors at other times, such that the commonness can be captured by our behavior relevance metric. Such an observation implies that the most recent month may be an informative period for depression diagnosis. Our method can potentially be leveraged, in a sliding window mechanism with the window size as the recent month, for early identification.

In addition to the number of days of data, the number of individuals used to build a good model also has important implications. 60 appears to be an inflection point in Figure 4. This suggests that such a user population size may be able to capture the majority of the behavior relevance for depression detection. However, 60 is not a plateau point. The performance still continues to have small improvements as the number of users increases. In our dataset, the results do not suggest a plateau yet, which indicates that a repository close to one-hundred users is still not large enough to capture all behavior relevance, and the performance can be further improved if more individuals' data is included. Our findings can provide a reference sample size for other researchers in the area.

6.3 Beyond Depression Detection

In addition to depression detection as a case study in this paper, our method has the potential to be applied on other daily-behavior-related binary classification tasks.

For example, in our institution 1 dataset, we have a small portion of students that had different BDI-II outcomes before and after the semester (26 from no depression to at least mild depression and 1 from mild to no depression). Detection of the change of depressive symptom can be an interesting and important task that worth investigation. A simple classification task is to mark users as being changed and unchanged, and then apply our method directly. It remains as an open question about how to incorporate the label dynamics. More generally, there are other tasks beyond depressive symptoms, such as the problem of loneliness [45] and the measure of mindfulness [11], that can also be explored using the same idea of collaborative filtering. When applying the method on a new classification task, the algorithms don't need to be changed. The behavior relevance metric still remains the same as it is only related to users' behavior trace rather than the specific classification goal. However, the feature selection results (Algorithm 2) and personalized interpretation outcomes (Algorithm 4) are depended on the classification target, thus leading to different results that are specific to a new task.

6.4 Limitations and Future Work

There are a few limitations in this work, which point out directions for future work. First, both datasets we used only have a post-semester/quarter BDI-II score to use as the ground-truth. Compared to some other work that has multiple labels for each individual (*e.g.*, [60]), we could not investigate more fine-grained individual dynamics of students' state other than by comparing their behavior relevance, which also hinders the investigation of the change of depression status, as mentioned in Section 6.3. In the future, we plan to collect ground-truth more frequently and apply our method on datasets with multiple labels for each individual to exploit individual dynamics. Having more labels will enable our algorithm to measure behavior relevance at a higher temporal resolution, explicitly take the depression dynamics into account, and produce more fine-grained predictions. We also plan to try semi-supervised learning, treating every two-weeks or every month (before the end of the semester/quarter) as unlabeled data.

Second, Algorithm 1 and Algorithm 2 treat each feature independently. It is the majority voting that integrates features across multiple streams and epochs. Also, our behavior relevance metric only requires the data to be aligned day by day. The temporal order of the data is not leveraged by the algorithm, *i.e.*, the classification results will not change if all users' data streams were re-sorted with the same order. We plan to work on new methods that can compile multi-modal features effectively and include the temporal aspect of our data in our current behavior metric. In addition, our method relies on the behavior features extracted from the dataset. Therefore, the capability of our method is limited by these features. There may exist more meaningful features to be discovered in the feature extraction part of our approach (Section 4.2), such as sleep hygiene patterns before bed and heart rate patterns, which may enable our algorithm to better capture behavior relevance among individuals. Moreover, there may exist better relevance metrics than square of correlation. For example, De Domenico *et al.*

[15] proposed to leverage mutual information to quantify the correlation between two multivariate nonlinear time series (mobility traces). There are more techniques to be explored in the future.

Third, we removed all participants who used Android phones because of the feature discrepancy across platforms (Section 4.1.2). We plan to investigate the platform distinction more deeply in future work. In addition, the removal of users who were missing more than half of their data might neglect some important cases: a day of missing data could be related to students' depression status (*e.g.*, not charging one's phone because of a diminished desire for social interaction [5]). It is almost impossible to distinguish whether the missingness is at random (due to various hardware or software issues) or caused by certain health-related factors. This could introduce both sampling bias and imputation bias. Compared to other simple imputation methods such as mean value imputation or previous-day imputation, our method better leverages population data. However, this method would inevitably introduce bias because the missing completely at random (MCAR) assumption is violated. It remains as an open question for how to incorporate participants with low-compliance appropriately.

7 CONCLUSION

In this paper, we present a new method for personalized behavior classification. Our method borrows the idea from memory-based collaborative filtering and uses the behavior correlation square to capture the behavior relevance among individuals. Moreover, it combines the relevance metric and association rule mining to obtain personalized behavior rules. We applied our method on a passive sensing dataset collected from 97 undergraduate students over one 16-week semester, whose depressive symptoms at the end of the semester were measured by their post-semester BDI-II scores. The results show that our method outperforms the state-of-the-art model on depression detection by 5.1% on the accuracy and 5.5% on the F1 score, with statistical significance. We further evaluated our method by replicating the pipeline on a second dataset collected from another institution and obtained similar results (3.4% improvement on average), demonstrating the pipeline-level generalizability of our method. The results also imply that more future work is needed to achieve model-level generalizability. Moreover, our method also generates highly interpretable rules that capture both the homogeneity and the heterogeneity in students' behavior related to depression, which could potentially inspire personalized depression intervention in the future.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grant Number IIS7974751 and EDA-2009977, the National Institute on Disability, Independent Living and Rehabilitation Research under Grant Number 90DPGE0003-01, Carnegie Mellon University Office of the Provost, University of Washington College of Engineering, Department of Electrical Engineering, and School of Computer Science & Engineering, Samsung Research America, and Google.

REFERENCES

[1] Saeed Abdullah, Nicholas D Lane, and Tanzeem Choudhury. 2012. Towards population scale activity recognition: A framework for handling data diversity. In *Twenty-Sixth AAAI Conference on Artificial Intelligence.*

[2] ACHA-NCHA II 2019. undergraduate student reference group - data report. https://www.acha.org/documents/ncha/NCHA-II_Spring_2019_Undergraduate_Reference_Group_Data_Report.pdf.

[3] 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.

[4] Karan Ahuja, Dohyun Kim, Franceska Xhakaj, Virag Varga, Anne Xie, Stanley Zhang, Jay Eric Townsend, Chris Harrison, Amy Ogan, and Yuvraj Agarwal. 2019. EduSense: Practical classroom sensing at scale. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 3, Article 71 (Sep 2019), 26 pages. https://doi.org/10.1145/3351229

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

[6] Christopher G Atkeson, Andrew W Moore, and Stefan Schaal. 1997. Locally weighted learning. In *Lazy Learning*. Springer, 11–73.

[7] 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.

[8] Sangwon Bae, Denzil Ferreira, Brian Suffoletto, Juan C. Puyana, Ryan Kurtz, Tammy Chung, and Anind K. Dey. 2017. Detecting drinking episodes in young adults using smartphone-based sensors. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 2, Article 5 (Jun 2017), 36 pages. https://doi.org/10.1145/3090051

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

[10] 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.

[11] Kirk Warren Brown and Richard M Ryan. 2003. The benefits of being present: Mindfulness and its role in psychological well-being. *Journal of personality and social psychology* 84, 4 (2003), 822.

[12] 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.

[13] 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).

[14] Taylor H Cox, Sharon A Lobel, and Poppy Lauretta McLeod. 1991. Effects of ethnic group cultural differences on cooperative and competitive behavior on a group task. *Academy of management journal* 34, 4 (1991), 827–847.

[15] Manlio De Domenico, Antonio Lima, and Mirco Musolesi. 2013. interdependence and predictability of human mobility and social interactions. *Pervasive and Mobile Computing* 9, 6 (Dec 2013), 798–807. https://doi.org/10.1016/j.pmcj.2013.07.008 arXiv:1210.2376

[16] 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.

[17] 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.

[18] Afsaneh Doryab, Daniella K Villalba, Prerna Chikersal, Janine M Dutcher, Michael Tumminia, Xinwen Liu, Sheldon Cohen, Kasey Creswell, Jennifer Mankoff, John D Creswell, and Anind K Dey. 2019. Identifying behavioral phenotypes of loneliness and social isolation with passive sensing: Statistical analysis, data mining and machine learning of smartphone and fitbit data. *JMIR Mhealth Uhealth* 7, 7 (24 Jul 2019), e13209. https://doi.org/10.2196/13209

[19] 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.

[20] 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.

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

[22] Jerome Friedman, Ron Kohavi, and Yeogirl Yun. 1997. Lazy decision trees. *Proceedings of the AAAI* 1 (09 1997).

[23] 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.

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

[25] Gabriella M Harari, Sandrine R Müller, Min SH Aung, and Peter J Rentfrow. 2017. Smartphone sensing methods for studying behavior in everyday life. *Current Opinion in Behavioral Sciences* 18 (2017), 83–90.

[26] Jin-Hyuk Hong, Julian Ramos, and Anind K Dey. 2015. Toward personalized activity recognition systems with a semipopulation approach. *IEEE Transactions on Human-Machine Systems* 46, 1 (2015), 101–112.

[27] Rob J Hyndman and Yanan Fan. 1996. Sample quantiles in statistical packages. *The American Statistician* 50, 4 (1996), 361–365.

[28] Wenjun Jiang, Qi Li, Lu Su, Chenglin Miao, Quanquan Gu, and Wenyao Xu. 2018. Towards personalized learning in mobile sensing systems. In *2018 IEEE 38th International Conference on Distributed Computing Systems (ICDCS)*. IEEE, 321–333.

[29] 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.

[30] James M Keller, Michael R Gray, and James A Givens. 1985. A fuzzy k-nearest neighbor algorithm. *IEEE Transactions on Systems, Man, and Cybernetics* 4 (1985), 580–585.

[31] Kurt Kroenke and Robert L Spitzer. 2002. The phq-9: A new depression diagnostic and severity measure. *Psychiatric annals* 32, 9 (2002), 509–515.

[32] Kurt Kroenke, Robert L Spitzer, Janet BW Williams, and Bernd Löwe. 2009. An ultra-brief screening scale for anxiety and depression: The phq–4. *Psychosomatics* 50, 6 (2009), 613–621.

[33] 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).

[34] Nicholas D. Lane, Ye Xu, Hong Lu, Shaohan Hu, Tanzeem Choudhury, Andrew T. Campbell, and Feng Zhao. 2011. Enabling large-scale human activity inference on smartphones using community similarity networks (csn). In *Proceedings of the 13th International Conference on Ubiquitous Computing* (Beijing, China) *(UbiComp '11)*. Association for Computing Machinery, New York, NY, USA, 355–364. https://doi.org/10.1145/2030112.2030160

[35] Ben Lengerich, Bryon Aragam, and Eric P Xing. 2019. Learning sample-specific models with low-rank personalized regression. In *Advances in Neural Information Processing Systems*. 3570–3580.

[36] Daniel Lopez-Martinez, Ognjen Rudovic, and Rosalind Picard. 2017. Physiological and behavioral profiling for nociceptive pain estimation using personalized multitask learning. *Neural Information Processing Systems Workshop on Machine Learning for Health* (2017).

[37] Jin Lu, Chao Shang, Chaoqun Yue, Reynaldo Morillo, Shweta Ware, Jayesh Kamath, Athanasios Bamis, Alexander Russell, Bing Wang, and Jinbo Bi. 2018. Joint modeling of heterogeneous sensing data for depression assessment via multi-task learning. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 2, 1, Article 21 (Mar 2018), 21 pages. https://doi.org/10.1145/3191753

[38] Stephen M. Mattingly, Julie M. Gregg, Pino G. Audia, Ayse Elvan Bayraktaroglu, Andrew T. Campbell, Nitesh V. Chawla, Vedant Das Swain, Munmun De Choudhury, Sidney K. D'Mello, Anind K. Dey, Ge Gao, Krithika Jagannath, Kaifeng Jiang, Suwen Lin, Qiang Liu, Gloria Mark, Gonzalo J. Martínez, Kizito Masaba, Shayan Mirjafari, Edward Moskal, Raghu Mulukutla, Kari Nies, Manikanta D. Reddy, Pablo Robles-Granda, Koustuv Saha, Anusha Sirigiri, and Aaron Striegel. 2019. The tesserae project: Large-scale, longitudinal, *in situ*, multimodal sensing of information workers. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems, CHI 2019, Glasgow, Scotland, UK, May 04-09, 2019*, Regan L. Mandryk, Stephen A. Brewster, Mark Hancock, Geraldine Fitzpatrick, Anna L. Cox, Vassilis Kostakos, and Mark Perry (Eds.). ACM. https://doi.org/10.1145/3290607.3299041

[39] Jun-Ki Min, Afsaneh Doryab, Jason Wiese, Shahriyar Amini, John Zimmerman, and Jason I. Hong. 2014. Toss "n" turn: Smartphone as sleep and sleep quality detector. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto, Ontario, Canada) *(CHI '14)*. Association for Computing Machinery, New York, NY, USA, 477–486. https://doi.org/10.1145/2556288.2557220

[40] Shayan Mirjafari, Kizito Masaba, Ted Grover, Weichen Wang, Pino G. Audia, Andrew T. Campbell, Nitesh V. Chawla, Vedant Das Swain, Munmun De Choudhury, Anind K. Dey, Sidney K. D'Mello, Ge Gao, Julie M. Gregg, Krithika Jagannath, Kaifeng Jiang, Suwen Lin, Qiang Liu, Gloria Mark, Gonzalo J. Martínez, Stephen M. Mattingly, Edward Moskal, Raghu Mulukutla, Subigya Nepal, Kari Nies, Manikanta D. Reddy, Pablo Robles-Granda, Koustuv Saha, Anusha Sirigiri, and Aaron Striegel. 2019. Differentiating higher and lower job performers in the workplace using mobile sensing. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 2 (2019), 37:1–37:24. https://doi.org/10.1145/3328908

[41] Robert H Moorman and Gerald L Blakely. 1995. Individualism-collectivism as an individual difference predictor of organizational citizenship behavior. *Journal of organizational behavior* 16, 2 (1995), 127–142.

[42] Mei Yi Ng and John R Weisz. 2016. Annual research review: Building a science of personalized intervention for youth mental health. *Journal of Child Psychology and Psychiatry* 57, 3 (2016), 216–236.

[43] NSDUH 2018. the national survey on drug use and health - survey report in 2018. https://www.samhsa.gov/data/sites/default/files/cbhsq-reports/NSDUHNationalFindingsReport2018/NSDUHNationalFindingsReport2018.pdf.

[44] Ognjen Rudovic, Jaeryoung Lee, Miles Dai, Björn Schuller, and Rosalind W Picard. 2018. Personalized machine learning for robot perception of affect and engagement in autism therapy. *Science Robotics* 3, 19 (2018), eaao6760.

[45] Daniel W Russell. 1996. Ucla loneliness scale (version 3): Reliability, validity, and factor structure. *Journal of personality assessment* 66, 1 (1996), 20–40.

[46] 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.

[47] 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).

[48] Tara N Sainath, Oriol Vinyals, Andrew Senior, and Haşim Sak. 2015. Convolutional, long short-term memory, fully connected deep neural networks. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 4580–4584.

[49] Peter Schulam and Suchi Saria. 2015. A framework for individualizing predictions of disease trajectories by exploiting multi-resolution structure. In *Advances in Neural Information Processing Systems*. 748–756.

[50] Yasaman S. Sefidgar, Woosuk Seo, Kevin S. Kuehn, Tim Althoff, Anne Browning, Eve Riskin, Paula S. Nurius, Anind K. Dey, and Jennifer Mankoff. 2019. Passively-sensed behavioral correlates of discrimination events in college students. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 114 (Nov 2019), 29 pages. https://doi.org/10.1145/3359216

[51] 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.

[52] Xu Sun, Hisashi Kashima, and Naonori Ueda. 2012. Large-scale personalized human activity recognition using online multitask learning. *IEEE Transactions on Knowledge and Data Engineering* 25, 11 (2012), 2551–2563.

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

[54] 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.

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

[56] Shyam Visweswaran and Gregory F Cooper. 2010. Learning instance-specific predictive models. *Journal of Machine Learning Research* 11, Dec (2010), 3333–3369.

[57] 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.

[58] 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

[59] 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.

[60] 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 (Mar 2018), 26 pages. https://doi.org/10.1145/3191775

[61] 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.

[62] Xuhai Xu, Prerna Chikersal, Afsaneh Doryab, Daniella K. Villalba, Janine M. Dutcher, Michael J. Tumminia, Tim Althoff, Sheldon Cohen, Kasey G. Creswell, J. David Creswell, Jennifer Mankoff, and Anind K. Dey. 2019. Leveraging routine behavior and contextually-filtered features for depression detection among college students. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 3 (2019), 116:1–116:33. https://doi.org/10.1145/3351274

[63] Makoto Yamada, Takeuchi Koh, Tomoharu Iwata, John Shawe-Taylor, and Samuel Kaski. 2017. localized lasso for high-dimensional regression. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (Proceedings of Machine Learning Research)*, Aarti Singh and Jerry Zhu (Eds.), Vol. 54. PMLR, Fort Lauderdale, FL, USA, 325–333. http://proceedings.mlr.press/v54/yamada17a.html

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

[65] Zhongtang Zhao, Yiqiang Chen, Junfa Liu, Zhiqi Shen, and Mingjie Liu. 2011. Cross-people mobile-phone based activity recognition. In *Twenty-second International Joint Conference on Artificial Intelligence*.

[66] Zijian Zheng and Geoffrey I Webb. 2000. Lazy learning of bayesian rules. *Machine Learning* 41, 1 (2000), 53–84.