Predicting The Severity of Anxiety in Adolescents Through Passively Sensed Behaviors

A Technical Report submitted to the Department of Computer Science

Presented to the Faculty of the School of Engineering and Applied Science

University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements for the Degree

Bachelor of Science, School of Engineering

Elaine Zhang

Spring, 2025

Technical Project Team Members

Matthew Clark

Afsaneh Doryab

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

Afsaneh Doryab, Department of Systems and Computer Science

Predicting The Severity of Anxiety in Adolescents Through Passively Sensed Behaviors

Elaine Zhang*
University of Virginia
Charlottesville, Virginia, USA
zzb2rf@virginia.edu

Matthew Clark*
University of Virginia
Charlottesville, Virginia, USA
wdd6ts@virginia.edu

Afsaneh Doryab
University of Virginia
Charlottesville, Virginia, USA
ad4ks@virginia.edu

Abstract

In this work, we investigate whether the severity of adolescents' anxiety can be predicted using passively sensed behaviors. We recruited 55 adolescent participants with diagnosed anxiety to participate in a 24-week-long study. Participants completed a weekly questionnaire recording their anxiety level. Using Fitbits and participants' mobile phones, we passively collected their physiology and behaviors. We find that adolescent anxiety can be predicted with a relatively low mean absolute error of 2.45 on a 22-point scale. Using SHAP values and feature importance, we determine that certain behaviors, particularly those related to phone usage, mobility, and activities, are particularly useful for predicting anxiety. Coincidentally, our study overlaps with the early stages of the COVID-19 pandemic. As such, we also explore how our participant's anxiety and behaviors varied throughout government-mandated lockdowns and during spikes of COVID cases after the lockdowns ended. We find evidence that participants were least anxious during partial lockdowns and when the cases of COVID-19 were low. We also find that participants' physiology and behaviors altered based on lockdown severity and prevalence of the disease. Despite these differences, the accuracy of our predictions remained consistent, regardless of lockdowns and the number of new COVID cases. These findings may provide insight into adolescents' anxiety and behaviors during prolonged traumatic events.

CCS Concepts

• **Human-centered computing → Empirical studies in ubiquitous and mobile computing**; **Smartphones**; • **Social and professional topics → People with disabilities**.

Keywords

Behavior Modeling, Anxiety, Passive Sensing, Machine Learning, COVID-19

ACM Reference Format:
Elaine Zhang, Matthew Clark, and Afsaneh Doryab. 2025. Predicting The Severity of Anxiety in Adolescents Through Passively Sensed Behaviors. In *ACM/IEEE International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE '25), June 24–26, 2025, New York, NY, USA*. ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3721201.3721389

*Both authors contributed equally to this work.

CHASE '25, June 24–26, 2025, New York, NY, USA

ACM ISBN 978-1-4503-14809.24/18/06
https://doi.org/10.1145/3721201.3721389

1 Introduction

According to the World Health Organization, anxiety disorders are the most common mental health problems in the world, affecting 301 million people in 2019 [47]. Despite its prevalence, only a quarter of people with anxiety receive adequate treatment. Anxiety in adolescents is particularly vital to study, as it can cause severe damage to children's self-esteem and mental well-being, possibly leading to withdrawn behavior and avoiding situations that make them anxious [34]. In adolescents, these behaviors may affect vital stages of cognitive, physical, and emotional development, resulting in detrimental mindsets and behaviors throughout their adult lives.

Despite the increasing prevalence of anxiety among teenagers and its effect on behavior, no study has investigated whether the severity of adolescents' anxiety can be predicted using passively sensed biobehavioral data from smartphones and wearable devices. Furthermore, there is a general lack of knowledge regarding the in-the-wild behaviors and physiology that are most indicative of an adolescent's anxiety. Investigating these gaps may help future mobile health systems to provide adolescents with personalized care for their anxiety without active and potentially costly assistance from mental health professionals.

In this work, we investigate whether the severity of anxiety in adolescents can be predicted using behaviors passively sensed via their smartphones and Fitbit devices. During a 24-week study, we collected smartphone and Fitbit data from 55 adolescents diagnosed with anxiety. Participants also completed a weekly GAD-7 Anxiety questionnaire. We extracted interpretable behavioral features summarizing the sensor data and used these features to predict participants' GAD-7 scores. Using feature importance and SHAP values, we also identify how specific behaviors, such as texting others, affect the prediction of these GAD-7 scores. Our analysis also investigated how prolonged traumatic events affect adolescent anxiety and behaviors. In particular, we focus on the influences of the COVID-19 pandemic, as our data collection began in early March 2020, less than two weeks before the government-mandated lockdown began.

The primary contributions of this work are:

(1) A novel investigation into how accurately the severity of *adolescents'* anxiety and its distinct characteristics can be predicted using behavioral features extracted from their mobile phones and wearable devices. We found that while classifiers failed to achieve acceptable accuracy, regressor models achieved low error rates.
(2) Insight into how different passively sensed behaviors contribute to predicting anxiety and its symptoms. We identified that features related to participants' location, screen time, and activities were the most useful to the models.

(3) Insight into how adolescent anxiety and behaviors varied throughout the COVID-19 pandemic. Our results show that participants were least anxious during partial lockdowns and less fearful when COVID was less prevalent. Our data also indicated significant differences in participants' movement, sleep, socialization, screentime, and heart rates between various pandemic stages.

2 Related Work

2.1 Sensing for Mobile Health

Mobile health technologies have greatly improved users' ability to gain health insights and suggestions without active intervention from clinicians. This field has benefited from the widespread adoption of mobile phones [38] and wearable devices [35]. These devices enable passive yet continuous monitoring of user activities and physiological states, leading to improved quality of patient care [48]. Mobile systems track a wide range of health concerns, such as sleep, mood, exercise, cardiovascular health, and eating habits [7, 12, 17, 27]. Although these apps and devices have become widespread, collecting meaningful health information without burdening users remains a challenge [33].

This challenge is highly prevalent in mobile health systems designed for mental health. Mental health diseases do not always manifest physical symptoms, making them difficult to monitor passively. Because of this challenge, mobile health-oriented apps tend to only provide general interventions, such as mindfulness training, rather than sensing the in-the-moment need for mobile health interventions [9]. More rigorous mobile health systems could significantly improve the clinical care of mental health patients [42]. This efficacy has been demonstrated, as mobile health systems have successfully demonstrated that the severity of ones' depression can be predicted using passively sensed behaviors [10]. However, apps that continuously track user behaviors tend to cause privacy and battery concerns [4]. As such, other systems have demonstrated that the severity of depression can be predicted through other behaviors, such as writing analysis [6, 49]. Although these approaches may mitigate privacy and battery concerns, they increase the use burden by requiring active engagement with the system. Most studies on mental health detection primarily focus on adults, often overlooking adolescents.

2.2 Mobile Healthcare for Adolescent Mental Health

Mobile health interventions targeting adolescents often neglect to collect data and preferences from the user, drastically limiting the ability to deliver personalized care [26]. In a sample of 121 commercial apps that help adolescent anxiety, zero passively collect behaviors, and more than 80% fail to even collect self-reported anxiety levels [5]. It has even been claimed that existing mobile health approaches for improving mental health have failed to demonstrate the ability to help adolescents [18].

As evidenced by mobile health systems for adults, enabling systems to glean health insights through passive sensing can improve their efficacy. However, very few studies have investigated the connections between mental health and in-the-wild behavior of adolescents. In the work most related to our own, Mullick et al. demonstrated that the severity of adolescent depression can be predicted through sensed behavioral features [32]. Several other studies have investigated the connections between adolescent mental health and passively sensed behaviors, but tend to either focus on a single behavior (such as screen time) [8, 19] or simply identify differences in behavior when participants are stressed, rather than predicting anxiety levels [29, 31]. Our work is the first to explore predicting the severity of adolescent anxiety using a comprehensive set of passively sensed behaviors from multiple sensor modalities.

2.3 Mental Health in Adolescents During the COVID-19 Pandemic

It is well documented that the COVID-19 pandemic had significant adverse effects on adolescent mental health [22, 23, 36, 37]. In a survey sent to 7,705 high school students in the United States, 37.1% of respondents reported experiencing poor mental health during the pandemic, a 6% increase compared to before the pandemic [23]. In particular, studies found that rates of anxiety, depression, ADHD symptoms, eating disorders, and general irritability all increased among adolescents during the pandemic [37]. Frequently, these studies seek to identify generalized personality and demographic risk factors for deteriorated mental health. The studies generally indicated that adolescents who prefer not to be at home [36], have weaker social connections [23], experience digital schooling [20], are older, low-income, or ethnic minorities [21] all experienced increased rates of mental distress.

Relatively few studies on mental health during COVID collected longitudinal data [37]. Moreover, these studies often relied on survey reports, which provide detailed descriptions of participants' mental states but frequently overlook how participant behavior impacts their mental health during the pandemic. [3, 13, 16, 30]. However, a few studies collect and analyze participant behavior in their surveys. A study among adolescents in France identified a decrease in naps and an increase in sleep disorders during the pandemic lockdowns [25]. Similarly, adolescents who reported engaging in frequent physical activity during the pandemic also reported lower anxiety levels [2]. Another study found that increased screen usage during the pandemic had no significant relationship with their emotional or cognitive states [24]. Although these studies provide insights into how various behaviors relate to adolescent anxiety, they rely entirely on self-reported data, which is vulnerable to response bias. Additionally, each study focuses on a single in-the-wild behavior, neglecting to comprehensively consider the nuances between multiple types of human behavior. To the best of our knowledge, our study is the first to analyze routinely collected anxiety reports and passively sensed behaviors from adolescents during the initial months of the COVID-19 pandemic.

3 Methods

3.1 Data Collection

Our study aimed to connect adolescent behavior in natural settings with their anxiety levels. It was designed to be minimally intrusive to ensure participants' behavior remained natural.

Participants were recruited from psychiatric clinics at a psychiatric hospital in the Mid-Atlantic region of the United States. Prospective participants first completed a screening assessment to ensure they met the criteria for inclusion in the study. To qualify, applicants needed to exhibit some anxiety symptoms, own a smartphone and be between 12 and 17 years old. Once these criteria were met, participants were asked to provide their demographic information through a RedCap survey. Finally, participants were instructed to install the Aware app [14] and were provided with a Fitbit Inspire HR [15], which facilitated the passive collection of their behaviors and physiological data.

Table 1: The GAD-7 questionnaire asks participants to report how frequently they experienced the emotions described, using a 4-point scale.

Question	Abbreviation
Feeling nervous, anxious, or on edge	NERV
Not being able to stop or control worrying	CTRL
Worrying too much about different things	WORRY
Trouble relaxing	RELAX
Being so restless that it is hard to sit still	RSTLS
Becoming easily annoyed or irritable	ANNOY
Feeling afraid, as if something awful might happen	AFRD

Out of 118 individuals who expressed interest in our study, only 55 participants completed the recruitment stage to be included in the data analysis. 41 participants self-identified as female, 12 as male, three as transgender, and two as non-binary. The average age of these participants was 15.527 years old ($\sigma = 1.512$). 51 participants identified as white, six as black or African American, three as Hispanic, one as Asian, and one as American Indian or Alaska Native. Finally, our participants reported a mean of 37 ($\sigma = 41.828$) trips to mental health professionals, and 83.636% had medical prescriptions for "emotional problems" at some time in their lives.

Participants had to keep the Aware app [14] running on their smartphones, which passively logs data from the devices' built-in sensors. The Aware app records data on screen time, location, phone calls, messages, WiFi & Bluetooth connections, and battery levels and uses the device's built-in accelerometer to detect activities. Samples from GPS, WiFi, and battery were recorded every 10 minutes. Data on phone calls, messages, and screen time were recorded whenever an event, such as a phone call, occurred. Similarly, participants were asked to wear the provided Fitbit whenever possible, allowing continuous recording of their steps, heart rate, and sleep.

Once a week, participants were required to complete the Generalized Anxiety Disorder 7-item scale (GAD-7) [43]. This questionnaire asks participants to describe their anxiety symptoms during the previous week. The GAD-7 consists of 7 questions, shown in table 1. For each question, participants report how often they were bothered by the problem described, using a four-point scale from zero to three. A score of zero means the problem has not bothered them at all, while a score of three represents they have experienced this symptom "Nearly every day." Higher scores on the questionnaire indicate that the participant experienced more severe anxiety. The GAD scores can also be categorized into four levels of anxiety using this questionnaire, as shown in Table 2. This questionnaire was deployed through RedCap.

Table 2: Anxiety levels and their respective GAD scores, as indicated by the GAD-7.

Anxiety Level	Category	Gad Scores
1	Minimal Anxiety	0 - 4
2	Mild Anxiety	5 - 9
3	Moderate Anxiety	10 - 14
4	Severe Anxiety	15 - 21

Each participant submitted data for 24 weeks, with data collection beginning in early March 2020. All participants were recruited before November 2020. Although our data collection overlaps with the COVID-19 pandemic, which is heavily analyzed in our results, this timing was coincidental.

3.2 Data Processing

3.2.1 Feature Extraction. We use RAPIDS [46] to extract weekly features from the time-series data collected by Aware and the Fitbit. The RAPIDS framework converts sensor readings to interpretable behavioral features. These features were extracted the week leading up to each GAD-7 response. This enabled us to build models predicting anxiety levels over a week, using the behaviors from that same week. We extracted 142 features to summarize each week. The feature set is briefly summarized in Table 3. These features describe a broad variety of behaviors including time spent at home, resting heart rate, and number of text messages sent. More details regarding RAPIDS can be found in [46].

Table 3: A summary of the 142 features extracted for our study. Calls, messages, locations, screen, WiFi, and activity were all recorded by the Aware app [14] on participants' phones. Sleep, steps, and heart rate were collected by participants' Fitbits [15].

Feature Group	Number of Features	Behavior Summarized
Calls	29	Phone calls made, received, and missed
Messages	10	SMS messages received and sent
Locations	34	Participants's movement
Screen	7	Participant's phone usage
WiFi	6	Connected and available WiFi points
Activity	6	Activity Recognition (stationary, walking, biking, etc.)
Sleep	36	Sleep timing and stages
Steps	5	Participant's steps taken
Heart Rate	9	Participant's heart rate

3.2.2 Handling Missing Values. Participants were removed from the dataset if more than 75% of their samples were missing. Missing samples resulted from non-compliance by either failing to complete the GAD-7, or not using the Fitbit or Aware app. This threshold of 75% was determined using the elbow method. This resulted in 29 patients being included in our final machine learning analysis.

Our dataset was also cleaned during the cross-validation process. In each cross-validation fold, any columns of a feature with all the data missing from the test or training set would be removed from

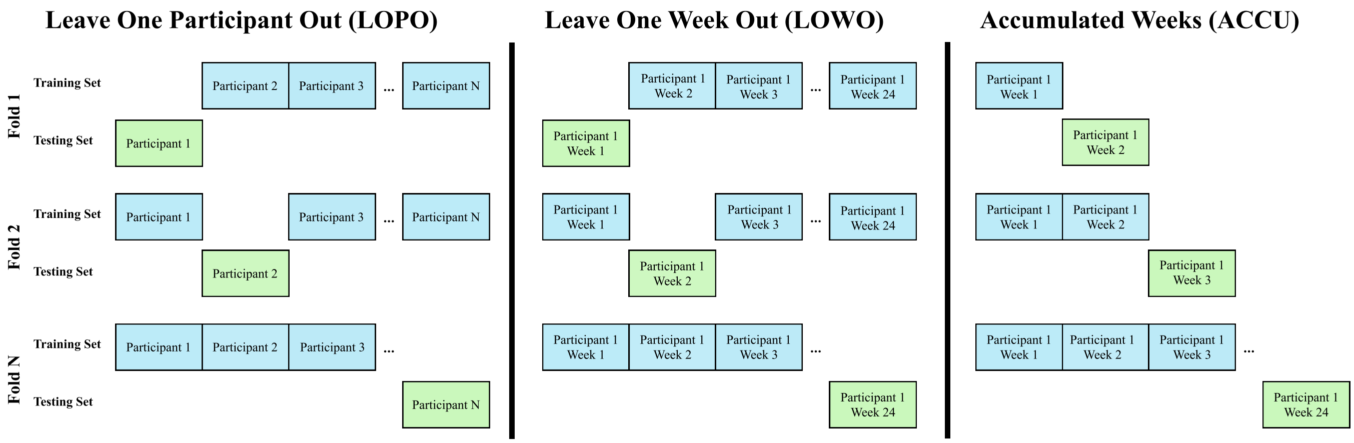

Figure 1: The three types of cross-validation used in our study. Leave One Week Out and Accumulated Weeks generate personalized models where each participant's data is trained and tested without considering any other participant's data.

both as it could not be imputed. If the feature was missing some, but not all, of the values in either the test or train set, we used feature imputation to infer missing values.

We imputed missing values by leveraging the time-series nature of our data. Our method exploits the assumption that weeks with the same anxiety levels would also have similar behavioral features. This assumption is based on established relationships between behavior and anxiety, such as anxiety disorders increasing the likelihood of insomnia, which would be reflected in our sleep features [45]. If a feature had a missing value in one week, we imputed the missing value using the values from adjacent weeks, as long as the weeks had the same anxiety level (Table 2). If a week had a missing feature, and its anxiety level was equal to the level for both the week before and after it, then the missing value was imputed as the mean of the known values from the adjacent weeks. If only one adjacent week had the same anxiety level, then the missing value was replaced with the known value from the adjacent week. If both adjacent weeks had different anxiety levels, the missing value was not imputed, and the feature was removed from the training and the test set. To prevent data leaks, the feature values in the test and training sets were never used to impute values in the other.

3.3 Statistical Analysis

Our statistical analysis focuses on identifying differences in participant behavior and anxiety throughout the COVID-19 Pandemic. This analysis primarily focuses on two external influences: the government-mandated COVID-19 lockdown and the rates at which COVID-19 was spreading. Unless otherwise stated, our statistical analysis was performed using a mixed linear model using Python's statsmodel package [41]. All of these models accounted for differences between participants.

COVID-19 Lockdowns. The state where our study was conducted implemented a three-level approach to COVID lockdowns. Each level was named after a color, with red indicating a complete lockdown and green indicating minimal restrictions. To analyze the effects of each lockdown stage, we compared the differences in behavioral features and GAD-7 responses collected during each lockdown level. The *red* level was from March 13 to May 15, 2020, where the state experienced complete lockdown [44]. During this phase, all non-essential stores and services were closed, with strict masking and social distancing regulations. All schooling during this phase was moved online. The *yellow* level of lockdown was from May 15 until June 5 and from July 3 until July 10, indicating a partial closure [39, 40]. During this phase, stores and restaurants could reopen at half capacity, and strict social distancing measures remained in place. During *green* level lockdown from June 5 to July 3 and after July 10, all businesses and services could reopen, as long as they continued to follow health guidelines [1].

COVID Spikes. We also tested how our participants' behaviors and anxiety varied in relation to the number of new COVID cases at the time the data was collected. The number of new COVID cases that occurred over each week in the area was calculated using the Johns Hopkins University COVID Data Repository [11] To reduce the influence of the lockdowns, this analysis only includes data collected after 10 July 2020, the end of the final lockdown. In addition to analyzing the differences across all participants, we consider that some participants may have been more conscientious about COVID's prevalence than others. To identify these individuals, we conducted Spearman R correlations between each participant's time at home and the number of new COVID cases over that week. We identified five participants who spent significantly more time at home while COVID spread faster. Due to their apparent awareness of COVID's prevalence, we sorted these five participants into a group we called "COVID-conscious". The remaining participants were grouped as "Typical Participants."

3.4 Machine Learning Analysis

Our models were trained to predict participants' responses to the GAD-7 questionnaire, computed anxiety levels, and each individual question in the GAD-7. The scores in this survey are ordinal. As such, we created both regression and classification models to predict these values. We used Adaboost, Decision Trees, Random Forests, Gradient Boosting, and XGBoost regressors and classifiers to predict

GAD-7 scores. We used three types of cross-validation as shown in Figure 1 and described below.

Leave One Participant Out (LOPO). This cross-validation simulated an environment where the system predicts the anxiety of an entirely unknown subject. In this cross-validation, one participant's data was removed as the test set, while all other participants' data was used as the training set. Each fold of the cross-validation used a different participant's data as the test set until all 29 participants' data had been tested.

Leave One Week Out (LOWO). This cross-validation created personalized models, where models train and test on a single participant's data. Each fold used a test set of a week's worth of data. The cross-validation repeated until every sample had been used as the test set. This process was completed for each participant's data, with accuracy aggregated across all cross-validations.

Accumulated Weeks (ACCU). This cross-validation simulated a scenario where the system gradually learns the subject's behavior. Each fold of the validation considered when the sample was collected. It only trained using samples collected before it. For example, when predicting anxiety levels during the third week of the study, the model was trained only on the sample collected during the first and second weeks. This form of cross-validation also resulted in personalized models. Previous work found that ACCU cross-validation resulted in the highest accuracy when predicting the Depression of adolescents [32].

Performance Metrics. To evaluate our regression models, we calculated the mean absolute error (MAE), mean squared error (MSR), and root mean squared error (RMSE). Our classification models were evaluated using accuracy, precision, recall, and F1-score. The metrics were recorded across every fold of the cross-validation and reported in aggregate.

Feature Importance. We also recorded the feature importance and SHAP values from each fold of our cross-validations. In particular, we sought to identify the features that were most frequently important to the models. To do so, we recorded the impurity-based feature importance for every feature in each fold of all cross-validations. For each fold, we calculated the median feature importance, identifying which features were more important than average. We then calculated the median number of models that considered each feature to be important. Features that had more of these important occurrences than the median were included as in the most frequently important feature set. This method enabled our analysis to identify the most frequently important features to the models.

4 Results

In this section, we investigated the following research questions:

RQ1: How accurately can adolescent anxiety be predicted through passively sensed behaviors?

RQ2: What behaviors were most important when predicting adolescent anxiety?

RQ3: How did the behavior and anxiety of our participants differ during government-mandated social-distancing measures?

RQ4: How did the prevalence of COVID-19 relate to the anxiety and behaviors of the participants after the official stay-at-home mandates were lifted?

4.1 Prediction of Anxiety in Adolescents

4.1.1 Adolescent Behaviors Can Be Used to Predict their Anxiety Scores. First, we investigated how well classification models can predict participants' GAD-7 scores. The best performing model was a Decision Tree using LOWO cross-validation, with an accuracy of 21%. Compared to the majority-class baseline, the accuracy of these models was consistently lower than if they merely guessed the most common label. This suggests that predicting the exact GAD-7 score using passively sensed behaviors may be unrealistic.

However, when observing the confusion matrix for these models, we noted that the predicted labels were often very close to the ground truth. As such, we repeated our analysis using regressors instead of classifiers. The results of this analysis are shown in Table 4. All of our models performed the worst in a LOPO cross-validation. The best LOPO model was the XGBoost regressor, with an MAE of 5.03. Interestingly, XGBoost was among the worst models in the LOWO and ACCU cross-validations. Our results during a LOWO cross-validation achieved the lowest average error, with Random Forest performing the best (MAE = 2.45). ACCU cross-validation resulted in error rates between LOPO and LOWO, with the Random Forest performing the best, achieving an MAE of 3.47. Since our models achieve mean absolute error rates as low as 2.45 on a 22-point scale, we can conclude these models were learning to predict anxiety. Given the low prediction accuracy by the classifiers but low error rates of the regressor models, future prediction of GAD-7 scores might be better approached as a regression problem, rather than classification.

4.1.2 Passively Sensed Adolescent Behaviors Can Be Used to Predict Anxiety Levels. The remainder of this analysis focuses exclusively on personalized models that performed best. The regression results of the LOWO and ACCU cross-validations are shown in Table 5. The best performing model was the Random Forest regressor using LOWO cross-validation (MAE = 0.49). For ACCU, the best performing model was also the Random Forest regressor (MAE = 0.87). Since the anxiety levels are on a scale of 4 points, we concluded that regression can accurately predict adolescents' anxiety levels.

In terms of classification, AdaBoost achieved the highest accuracy when using a LOWO cross-validation, with an accuracy of 0.58. This outperformed ACCU's most accurate model, Random Forest, achieving 0.55 accuracy. Figure 2 shows the confusion matrices associated with these models. When compared to the majority-class, for AdaBoost LOWO cross-validation for each patient, these models also generally predict below the the proportion of labels belonging to the majority class. This further supports our earlier finding that regressor models appear to perform better in predicting ordinal anxiety levels than classifiers.

4.1.3 Passively Sensed Adolescent Behaviors Can Be Used to Predict Certain Aspects of Anxiety. We then analyzed how accurately different aspects of anxiety could be predicted. We trained models to predict participants' responses to each question on the GAD-7, using the behavioral features. The results from the best perfoming

Table 4: The Accuracy of GAD-7 Score Predictions, across different regressors and cross-validations. A Random Forest with a Leave-One-Week-Out (LOWO) cross-validation achieved the lowest error.

	LOPO			LOWO			ACCU		
Model	**MAE**	**MSE**	**RMSE**	**MAE**	**MSE**	**RMSE**	**MAE**	**MSE**	**RMSE**
AdaBoost	5.2	40.47	5.85	2.5	12.31	2.5	3.57	28.59	3.57
Decision Trees	6.17	60.41	7.24	2.91	17.97	2.91	3.75	30.96	3.75
Random Forest	5.13	39.02	5.8	**2.45**	**10.74**	**2.45**	**3.47**	**25.02**	**3.47**
Gradient Boosting	5.29	42.36	5.95	2.71	13.95	2.71	3.59	27.53	3.59
XGBoost	**5.03**	**39.05**	**5.77**	2.78	16.1	2.78	3.72	30.52	3.72

Table 5: The Accuracy of Anxiety Level Predictions, across different regressors and cross validations. A Random Forest Leave-One-Week-Out (LOWO) cross-validation achieved the lowest error.

	LOWO			ACCU		
Model	**MAE**	**MSE**	**RMSE**	**MAE**	**MSE**	**RMSE**
AdaBoost	0.49	0.56	0.49	0.87	2.01	0.87
Decision Trees	0.46	0.75	0.56	0.9	2.13	0.9
Random Forest	**0.49**	**0.42**	**0.49**	**0.87**	**1.82**	**0.87**
Gradient Boosting	0.54	0.58	0.54	0.88	1.96	0.88
XGBoost	0.56	0.66	0.56	0.89	2.05	0.89

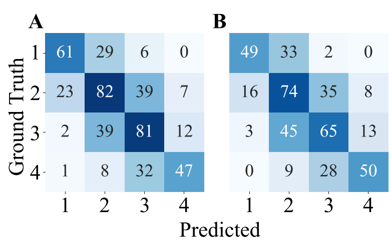

Figure 2: Confusion matrix results for predicting anxiety levels. A) Adaboost model with LOWO cross-validation. B) Random forest model with ACCU cross-validation.

regressors are shown in Table 6. The best regressor for each question was a Random Forest using a LOWO cross-validation. Most questions have a mean absolute error close to 0.5, except RELAX, which is closer to 0.57.

We also investigated how well classifiers could predict a patient's responses to each gad question. These results are shown in Figure 3. For all questions other than RELAX, the models achieved an accuracy between 55% and 63%. Accuracy for RELAX was slightly lower, at 52%. However, when analyzing the majority-class for the best-performing models by the patient, the model accuracy score per patient is lower than the majority-class baseline signifying that classification does not perform well in predicting anxiety levels. We again observe regression models perform better than classifiers when predicting anxiety levels. Interestingly, our regressors predicting RELAX achieved the highest error rates, as well as the lowest

Table 6: Regressor error rates for each question in the GAD-7. For each question, the best-performing model was a Random Forest using a LOWO cross-validation.

	LOWO			ACCU		
Question	**MAE**	**MSE**	**RMSE**	**MAE**	**MSE**	**RMSE**
NERV	0.51	0.42	0.51	0.73	1.18	0.73
CTRL	0.5	0.47	0.5	0.73	1.16	0.73
WORRY	0.5	0.43	0.5	0.76	1.3	0.76
RELAX	0.57	0.56	0.57	0.84	1.37	0.84
RSTLS	0.51	0.45	0.51	0.77	1.24	0.77
ANNOY	0.51	0.46	0.51	0.77	1.4	0.77
AFRD	0.5	0.48	0.5	0.72	1.17	0.72

accuracy, indicating a potential disconnect between responses to this question and participant behaviors.

4.2 Behavioral Features Predictive of Anxiety

Next, we explore what behavioral features were most predictive of anxiety. Our feature importance analysis focused on predicting GAD-7 scores, the most comprehensive of our anxiety measurements. We report the results from the best-performing model, a Random Forest Regressor, in a LOWO cross-validation.

4.2.1 Location, Screen, and Activity Features Were Most Commonly Important to the Models. We first identified the most frequently selected features when predicting GAD-7 scores and explored which types of features were most frequently selected during the cross-validation process. All seven of the phone screen features were frequently selected, highlighting a relationship between participants' anxiety and how they use their phones. Five of the six (83% of) activity recognition features were selected. All five relate to the amount of time spent doing different activities (running, stationary, biking, walking, and in a vehicle). Next, 25 of the 34 location features (73%) were selected, including all features that measured distance traveled, emphasizing that participants' mobility may vary based on their anxiety. 20 of the 29 (69%) call features were selected as frequently important, focusing mainly on the number, length, and timing of incoming and outgoing calls. Finally, 3 (50% of) WiFi, 4 (50% of) heart rate, and 4 (11% of) sleep features were identified as frequently selected. These results indicate that location, screen, and

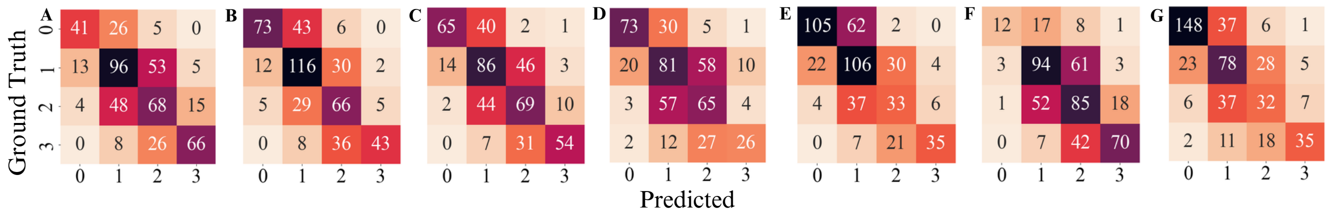

Figure 3: Confusion matrix for the predictions of each GAD-7 question. A) AdaBoost (AB) for NERV, 57% accuracy. B) Random Forest (RF) for CTRL, 63% accuracy. C) RF for WORRY, 58% accuracy. D) AB for RELAX, 52% accuracy. E) RF for RSTLS, 59% accuracy. F) RF for ANNOY, 55% accuracy. G) AB for AFRD, 62% accuracy.

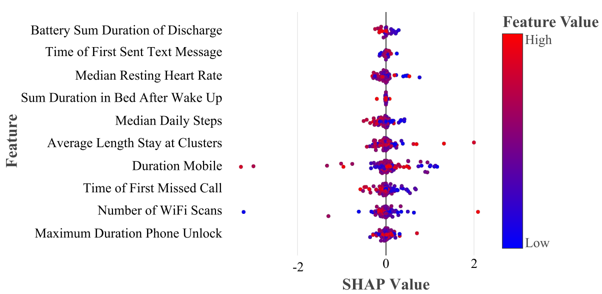

Figure 4: SHAP Values and Feature Values for one feature from each sensor. Positive SHAP values indicate the feature increased the model performance, while negative SHAP values indicate the feature increased the model's error. Feature values with a higher absolute value had a stronger influence on the model. According to Mann-Whitney U tests, each displayed features' values differ significantly between positive and negative SHAP values.

activity recognition features are more indicative of anxiety, while sleep is far less important than other behaviors.

4.2.2 Location, Sleep, and Socialization Behaviors Significantly Related to Model Accuracy. To further understand these models, we calculated the SHAP values from each cross-validation fold using Python's SHAP package [28]. SHAP values measure how a feature altered the final prediction of the final model. Positive SHAP values indicate that the feature increased the prediction, while negative SHAP values indicate the feature decreased it. SHAP values near 0 have minimal effect on the output, while features with high absolute values have more impact. The distribution of SHAP and feature values for the ten most frequently selected features are shown in Figure 4. Although this figure only shows ten features, SHAP values were extracted and analyzed for all 142 features.

For every feature, we used a Mann-Whitney U test to compare the feature value when the SHAP value is positive against the feature values when the SHAP is negative. This analysis revealed 45 features with significant differences. Most (62%) of these features indicated higher feature values made the models less accurate, while only 38% benefited from higher values. Below, we discuss the common themes within these features.

Eleven location features had significant relationships between feature values and the sign of the SHAP value. The feature quantifying variance in location ($U = 3.909, p = 1.352 * 10^{-4}$), the amount of time calling others ($U = 2.732, p = 0.006$), the time they spent mobile ($U = 4.593, p = 4.379 * 10^{-6}$), had a positive influence on model performance when the feature values were larger ($U = 3.909, p = 1.352 * 10^{-4}$), whereas higher values of the feature measuring the amount of time spent in bed after waking up had a negative influence on the model performance ($U = -3.999, p = 6.343 * 10^{-5}$). The negative influence was lower for larger feature values that measure the time spent awake ($U = 2.257, p = 0.024$) and napping ($U = 3.578, p = 3.463 * 10^{-4}$). The negative influence was also lower for smaller feature values that quantify the timing of calls ($U = -3.483, p = 4.965 * 10^{-4}$), the number of steps taken each day ($U = -4.679, p = 2.889 * 10^{-6}$), the number of text messages received ($U = -3.411, p = 6.465 * 10^{-4}$), and the maximum time a participant spent on the phone ($U = -4.041, p = 5.316 * 10^{-5}$). These trends provide interesting insight into how certain behavioral features may have impacted the accuracy of our models.

4.3 Adolescent Behavior and Anxiety During the COVID-19 Lockdown

We then analyzed how the government-mandated COVID-19 lockdowns potentially influenced our participants' anxiety and behaviors. During the red lockdown phase, we collected 120 GAD-7 surveys from 23 participants. 106 questionnaires were collected from 38 participants during level yellow lockdowns, and 864 responses were collected from 52 participants during the green phases of lockdown. We also collected 15 questionnaires from 9 participants before the COVID lockdowns started. However, these 15 responses were omitted from analysis due to the small sample size.

4.3.1 Participants Were Least Anxious During Phase "Yellow" Lockdowns. We analyzed differences in GAD-7 responses between the lockdown phases. The results of this pairwise analysis are shown in figure 5. Participants' GAD-7 scores were significantly lower during yellow level lockdowns compared to green ($p = 0.043, \beta = -0.805$), although there were no significant differences when comparing green to red ($p = 0.817, \beta = 0.103$) or red to yellow ($p = 0.074, \beta = -0.908$). Anxiety levels during the yellow lockdown phases were significantly lower than during the green phases ($p = 0.033, \beta = -0.169$) and red phase ($p = 0.021, \beta = 0.233$). There was no significant difference in anxiety levels between the red and green phases

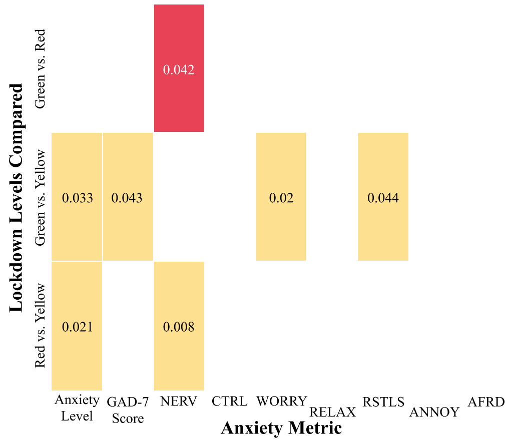

Figure 5: Differences in the anxiety metrics between the different lockdown phases. The color of the cell denotes which lockdown phase (red, green, or yellow) had lower values (less anxiety). The number in each box represents the p-value. Empty boxes indicate there was no significant difference.

($p = 0.471, \beta = -0.064$). These results indicate that the participants' anxiety was less severe during partial COVID lockdowns.

Participants' responses to the individual GAD-7 questions demonstrate a similar trend. NERV responses during the red phase were significantly lower than green ($p = 0.042, \beta = -0.180$) but higher than yellow ($p = 0.008, \beta = -0.267$). Participants' scores for WORRY were significantly lower during yellow phases than green ($p = 0.020, \beta = -0.196$), although there were no significant differences between yellow and red ($p = 0.117, \beta = 0.169$), or green and red ($p = 0.776, \beta = -0.027$). Lastly, responses to RSTLS were also significantly lower during yellow lockdown levels compared to green ($p = 0.044, \beta = -0.165$), but no significant differences were found between yellow and red ($p = 0.264, \beta = 0.117$), or red and green ($p = 0.603, \beta = 0.048$). We found no significant differences in responses for the CTRL, RELAX, ANNOY, and AFRD questions. These models further support the finding that participants were generally less anxious during yellow phase lockdowns when society was reopening.

4.3.2 Participants' Movement, Sleep, and Socialization Differed Between Lockdown Phases. Next, we analyzed how behaviors differed between the various lockdown phases. The features that varied significantly between lockdown levels are shown in Figure 6.

This analysis identified several behaviors collected by participants' Fitbits that differed between the lockdown phases. Two sleep features were present, indicating participants spent less time lying in bed awake during red-level lockdowns than green. Five features also significantly differed, indicating that participants generally took fewer daily steps during red lockdowns than green or yellow. Similarly, five heart rate features all indicated that participants' resting heart rates were lower during green-level lockdowns than yellow.

Several activity recognition and location features were identified as differing between phases. The activity recognition features indicated that participants spent less time in vehicles, moving, and being stationary during the red-level lockdown. Similarly, the 16 location features indicated that participants traveled significantly more during yellow and green lockdown levels than during red.

Our results also identified that participants' messaging and phone call behaviors differed between the lockdown stages. 14 phone call features were selected. These features indicate that participants made and received significantly more calls during yellow-level lockdowns. Similarly, the included messaging features indicate that participants sent and received fewer messages during the green phases of the lockdown.

We also identified differences in five features related to participants' screen time. Participants unlocked their phones significantly fewer times during green lockdowns than during yellow. Also, unlocks tended to be longer during red lockdowns than green or yellow ones. Notably, however, our results indicate that the total time participants spent on their phones did not vary between lockdown levels.

4.3.3 Predictions of Anxiety have Similar Error Rates Across Lockdown Phases. Finally, we analyzed whether the accuracy of our models differed between the lockdown phases. We identified that models that predicted the GAD-7 scores for weeks in the red phase had the lowest mean absolute error ($\mu = 2.298, \sigma = 1.790$). However, this performance was insignificantly lower than the predictions for the yellow ($\mu_{\text{yellow}} = 2.675, \sigma_{\text{yellow}} = 2.037, p = 0.592, \beta = -0.140$) and red periods ($\mu_{\text{red}} = 2.718, \sigma_{\text{yellow}} = 2.822, p = 0.761, \beta = 0.067$). Similarly, no significant difference was detected between the yellow or green lockdown phases ($p = 0.319, \beta = -0.207$). This indicated that adolescent anxiety remains equally possible to predict from behavioral features, regardless of lockdown levels.

4.4 Adolescent Behavior and Anxiety Compared to COVID Prevalence

We next analyzed how our participants' behavior and anxiety related to the number of new COVID cases at the time.

4.4.1 The Prevalence of COVID-19 Correlates with Participants' Fear, But Not Overarching Anxiety. We create models to identify relationships between participants' weekly GAD-7 scores and the number of new COVID cases over the same week. The number of new COVID cases was not significantly related to participants' GAD-7 scores ($p = 0.379, \beta = 4.780{*}10^{-5}$). Similarly, most questions within the GAD-7, r NERV ($p = 0.711, \beta = -4.440 * 10^{-6}$), CTRL ($p = 0.564, \beta = 6.599 * 10^{-6}$), WORRY ($p = 0.687, \beta = 4.757 * 10^{-6}$), RELAX ($p = 0.401, \beta = -1.013 * 10^{-5}$), RSTLS ($p = 0.182, \beta = -1.537^{-5}$), and ANNOY ($p = 0.703, \beta = 4.840 * 10{-}6$), did not have significant relationships with the number of new COVID cases. However, the final question of GAD-7, AFRD, had a significant positive relationship with the prevalence of COVID ($p = 2.258 * 10^{-5}, p = 3.471 * 10^{-2}$). This indicated while the severity of COVID did not relate to our participants' anxiety and most of its symptoms, they may have been more fearful when the disease was spreading more.

We also identified two differences in the responses from COVID-conscious and typical participants. First, COVID-conscious participants' responses to RSTLS had a significant negative relationship with the number of new COVID cases ($p = 0.028, \beta = -6.657{*}10^{-5}$),

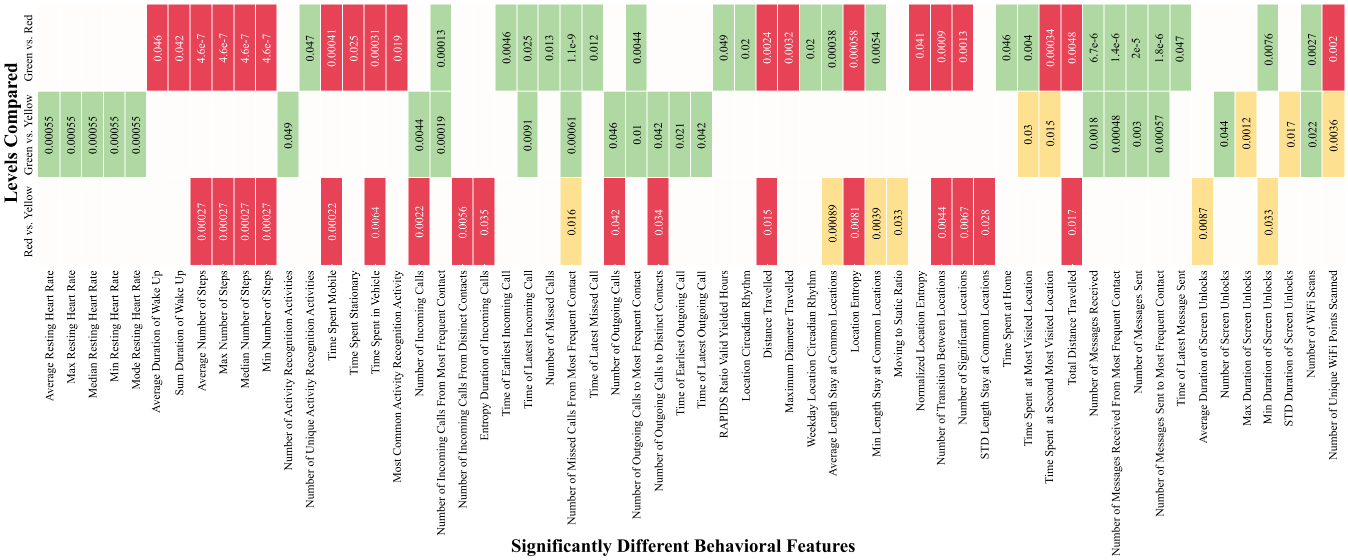

Figure 6: Behavioral features that significantly differed between the various lockdown levels. Each square contains the p-value, and the color denotes the lockdown level (red, green, or yellow) with lower values for that feature. Empty squares indicate there was no significant difference.

while typical participants did not ($p = 0.600, \beta = -6.533 * 10-6$). Similarly, responses to AFRD differed. The significant positive relationship identified above by the model using the data of all participants is reflected in the COVID-conscious participants ($p = 9.800 * 10^{-4}, \beta = 6.858 * 10^{-5}$), but not by the typical participants ($p = 0.223, \beta = 1.455 * 10^{-5}$). These results indicate that COVID-conscious participants might have been more fearful but less restless when COVID was more prevalent, but typical participants' anxiety may not have changed significantly.

4.4.2 COVID-Conscious and Typical Participants' Behaviors Were Similar, But Their Physiology Differed During COVID Spikes. We then analyzed the relationships between the behavioral features of the COVID-conscious and typical participants and COVID-19 prevalence at the time the behaviors were recorded. We discarded missing values for every feature and ensured every test had at least 30 feature samples.

Participants' location features generally followed similar trends regardless of whether they were COVID-conscious. As expected, COVID-conscious participants spent significantly more time at home when COVID was more prevalent ($p = 8.173 * 10^{-6}, \beta = 0.041$), while typical participants did not ($p = 0.068, \beta = 0.010$). Interestingly, typical participants traveled significantly less distances when there were more new COVID cases ($p = 0.010, \beta = -0.711$), while COVID-conscious participants did not ($p = 0.064, \beta = -0.9637$). Thus, while typical participants may not have stayed home more during COVID spikes, they did travel less, indicating they may have still taken safety precautions.

Participants were also more social on their phones when COVID was more prevalent. When there were more new COVID cases, both COVID-Conscious ($p = 1.138 * 10^{-4}, \beta = 0.066$) and typical ($p = 0.018, \beta = 0.056$) spent more time calling others. Interestingly, typical participants also started more calls ($p = 3.635 * 10^{-4}, \beta = 1.033 * 10^{-4}$), while COVID-conscious participants did not ($p = 0.651, \beta = -1.58 * 10^{-5}$). We also observed that typical participants sent ($p = 8.662 * 10^{-5}, \beta = 0.001$) and received ($p = 6.681 * 10^{-4}, \beta = 4.197^1 0^{-4}$) more text messages when COVID was more prominent. Not enough messaging data was collected from COVID-conscious participants accurately to analyze their behaviors. However, across all participants the number of messages sent ($p = 1.741 * 10^{-5}, \beta = 0.001$) and received ($p = 0.002, \beta = 4.46^1 0 * -4$) also identified a significant increase when COVID was more prominent. These trends are marginally weaker than those identified in typical participants only, indicating that there could be differences in messaging behaviors from the COVID-conscious participants. As such, our data indicates that COVID-conscious and typical participants both spent more time calling others when COVID was spreading more.

The physiology of the participants also contained two key differences between the groups. The average resting heart rate of the COVID-conscious participants was lower when COVID was more prevalent ($p = 0.002, \beta = -3.224 * 10^{-4}$), while typical participants' average resting heart rate was higher ($p = 0.031, \beta = 1.854 * 10^{-4}$). Furthermore, typical participants slept significantly more efficiently when COVID was more severe ($p = 0.017, \beta = 1.565 * 10^{-4}$), while COVID-conscious participants' data did not contain a significant trend ($p = 0.755, \beta = -2.124^{-5}$). This may indicate participants' physiology may have differed during COVID spikes.

4.4.3 Models Predicting Anxiety Achieved Similar Error Rates for COVID-Conscious and Typical Participants, but Used Slightly Different Features. Finally, we compared the models created to predict the anxiety of COVID-conscious and typical participants. We first investigated the mean absolute error for COVID-conscious ($\mu = 2.681, \sigma = 2.794$) and typical ($\mu = 2.674, \sigma = 2.668$) participants.

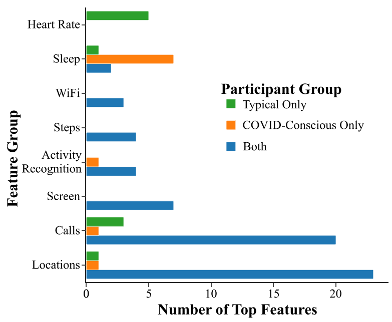

Figure 7: The most frequently important features among the COVID-Conscious and Typical participant groups. Despite many similarities, the models predicting anxiety for COVID-Conscious participants tend to use more sleep features, while those for typical participants tend to use heart rate features.

No significant difference was detected ($p = 0.909$), indicating that the models are likely equally accurate for both types of participants.

We repeated our feature importance analysis from section 3.4, separating the COVID-conscious and typical participant groups. First, we counted the number of top occurrences of each feature among each feature group. Using a Chi-Squared test, we identified a significant difference in the most frequently important aspects between the two groups ($\chi^2 = 1939.300, p < 0.001$).

To identify these differences in greater detail, we identified the most commonly important features for models belonging to COVID-conscious and typical participants. These results are shown in Figure 7. Most features are shared between the two feature groups. However, models predicting anxiety for COVID-conscious participants were more likely to use sleep features, while typical participants' models utilized more heart rate features. This may indicate differences in how participants' behaviors relate to their anxiety between the two groups.

5 Discussion

5.1 Mobile Sensing Data Can Be Used to Predict Different Metrics of Adolescent Anxiety

Our regression models predicted the anxiety level of 29 adolescent participants using passively sensed behaviors with a relatively low error rate. Although we achieve low error rates, it is challenging to directly compare our models' error to those of similar studies due to different anxiety and performance metrics. Even though our error is low for the 22-point GAD-7, model performance should still be improved to minimize errors when generalizing the predicted GAD-7 score to the more granular anxiety levels (Table 2). We also discovered that these models perform the best when trained and tested on data from a single participant, emphasizing the need for individualized models. Our results also indicated that a leave-one-week-out cross validation results in more accurate models than an accumulated-week approach. Additionally, our regressor models achieved a low error, but our classifiers failed to achieve acceptable accuracy. This may indicate that the exact anxiety level may be unrealistic to predict, but regressor models can provide a prediction with reasonable error.

Not only can our models predict anxiety scores, but they successfully predict specific aspects of anxiety, as defined by the GAD-7. Mobile health systems may leverage passively sensed behaviors to accurately predict other forms of cognitive and emotional distress, such as irritability, nervousness, and restlessness. Interestingly, the accuracy of these models was relatively consistent across anxiety aspects, with the notable exception of predicting when participants cannot relax, which had a higher error rate. Since physiological indicators like heart rate are generally expected to identify relaxed states, it is notable that our collected heart rate data did not result in a well-performing predictor of a relaxed state. Future work may investigate how relaxation can be accurately predicted through passively sensed behaviors.

5.2 Adolescent Location, Screen Time, and Activities are most Indicative of their Anxiety

Our analysis found that behaviors related to phone screen usage, location, and phone activity recognition were frequently important when predicting anxiety. In particular, these features demonstrate that GAD-7 scores were often predicted using mobility and phone usage features. Interestingly, sleep features were rarely important to the models, indicating adolescent sleep patterns are not as related to their anxiety as movement and phone usage.

The explainability of our models was further enhanced by our analysis of the models' SHAP values. We found location, sleep, calls, and messaging features had the most features significantly related to the model's performance. Interestingly, most of these significant relationships implied that high feature values would increase error, while lower values would decrease it. Future work can further investigate the differences between these features and SHAP values, better explaining why doing less of most behaviors appears to make anxiety easier to predict. Our analysis of location, and activity recognition features also indicated that the models were more accurate when participants moved more, but this was not reflected in the COVID analysis. We observed that our participants were less mobile during the Red lockdown phases, but the models remained equally accurate during all three levels of lockdown.

5.3 The Severity of Anxiety Changed Throughout the Pandemic

We analyzed participants' anxiety throughout the pandemic throughout the lockdown phases, and in relation to the number of new COVID cases. In particular, participants were less anxious during partial COVID lockdowns than during full or no lockdown. The improvement over complete lockdowns likely indicates participants were optimistic that the pandemic was ending and excited to be permitted to leave home again. These positive changes may have led to overall decreased anxiety. However, this change was not reflected during the comparison between partial and no lockdowns. This difference could indicate that the extra COVID precautions in place during a partial lockdown helped alleviate the anxiety of the participants, and once those precautions were lifted, their anxiety increased. This could also be due to an ordering effect, as participants could be optimistic about the restrictions easing during a partial lockdown. Still, the novelty could have worn off when the lockdowns were lifted. These findings somewhat contradict the existing literature that indicates that mental health decreased drastically during lockdowns and improved after the lockdowns were lifted.

We also found that once the lockdowns were fully lifted, participants' anxiety did not directly relate to the severity of COVID at that time. While GAD-7 scores did not differ, several individual characteristics did. In particular, participants were more fearful when the disease spread faster, and those taking COVID precautions were less restless. Despite reporting increased fearfulness, other GAD-7 questions, particularly those that quantified nervousness and worry, did not change. This could indicate that while many participants were monitoring the prevalence of COVID, and these changes increased their fear, they accepted that the disease was widespread and were not as worried about getting it. It also demonstrated that prolonged exposure to the pandemic did not make participants less afraid of the disease but did make them less fearful of getting it.

5.4 Adolescent Behaviors Varied Throughout the Pandemic Stages

To the best of our knowledge, our work is the first to study the passively sensed behaviors of adolescents during the COVID-19 pandemic. As expected, we observed many behaviors that varied between the various lockdown stages and between different participants. Although these differences are expected, the exact differences provide interesting insight into the behavior of adolescents.

First, we identified that participants generally spent less time lying in bed in the morning during full lockdowns than without lockdown. This finding is somewhat counterintuitive, as participants would be expected to spend more time lying in bed when they cannot leave home, and have less to do. However, our data indicated the opposite. Similarly, participants' location data demonstrated that participants did not spend more time at home during partial lockdowns compared to no lockdowns, despite government guidance to stay home whenever possible. This may indicate that our participants did not closely adhere to governmental guidance once the lockdowns were partially lifted. Finally, our participants utilized their phones to interact with others significantly less during no lockdown compared to full and partial lockdowns. Intuitively, our participants interacted with others face-to-face once restrictions were lifted, whereas they had to do so digitally when restrictions were active. We also found that participants spent more time at locations other than home during partial lockdowns, compared to no lockdowns. Yet, the communication data indicates they remained socially distant, not interacting with others in person. These findings provide insight into adolescent behavior during the pandemic, but more research is required to properly explain these behaviors beyond intuitive speculation.

5.5 Future Work & Further Limitations

Although much of our work focuses on behaviors and mental health during COVID-19, our study was not explicitly designed for this purpose. As a result, our analysis contains potential confounds that could not be prevented, such as a lack of random ordering between lockdown conditions. While our study contains interesting findings connecting adolescent behaviors to anxiety during the pandemic, these confounds are present throughout the analysis and are important to consider.

Missing data was a considerable limitation of this study. Only 29 of the 55 participants (53%) made it to the final machine learning dataset due to missing GAD-7 questionnaires or sensor data. Within the 29 included participants, there was still much missing data that had to be imputed with a moving average. This may be because our study collected a single questionnaire a week, and if participants failed to complete it, the entire week of data had to be discarded. Future studies could collect GAD-7 scores more frequently, allowing daily predictions rather than weekly, minimizing the effect of missing data. We also could not use many of the passively sensed Fitbit data in these models due to the lack of data. As a result, we miss key insights into whether or not sleep, heart, or step habits could be able to predict anxiety for specific participants.

Finally, our COVID analysis uses lockdown dates and prevalence data for a specific geographic area. We could not access their precise county of residence to preserve participant privacy. Since all participants were recruited in person, our analysis considered all counties near the hospital, but it is possible that some participants lived outside of this area.

6 Conclusion

We conducted a study to investigate whether the severity of adolescents' anxiety can be predicted through passively sensed behaviors. 55 participants were recruited to wear a Fitbit, passively record data through their phone, and complete weekly questionnaires quantifying their anxiety for 24 weeks. We determined that participants' responses to the anxiety questionnaire could be predicted with relatively low error rates. Using feature importance and SHAP values to interpret these models, we identify that features related to mobility and phone usage are the most useful for predicting anxiety. Coincidentally, our data collection began in early March 2020, enabling us to study the effects of the COVID-19 pandemic on participant behaviors and anxiety. We find that our anxiety predictions achieve similar accuracy, regardless of government-mandated lockdowns or the prevalence of the disease. Participants experienced lower anxiety levels during partial lockdowns and felt less fear when the disease was spreading at a slower rate. However, their overall anxiety did not directly correlate with the prevalence of the disease. Our findings can contribute to the development of mobile health systems that passively monitor adolescent anxiety, offering valuable insights into potential behavior and mental health during future pandemics.

Acknowledgments

We thank Ana Radovic and Sam Shaaban for their assistance with this manuscript and study. This work was supported by the National Institute of Mental Health (1R44MH122067).

References

[1] Reopening phase orders updated to include 10 additional counties moving to yellow and 16 to green on june 5. *Office of the Governor* (June 2020).

[2] Alves, J. M., Yunker, A. G., DeFendis, A., Xiang, A. H., and Page, K. A. Prenatal exposure to gestational diabetes is associated with anxiety and physical inactivity in children during covid -19. *Clinical Obesity 11*, 1 (Feb. 2021), e12422.

[3] Bignardi, G., Dalmaijer, E. S., Anwyl-Irvine, A. L., Smith, T. A., Siugzdaite, R., Uh, S., and Astle, D. E. Longitudinal increases in childhood depression symptoms during the covid-19 lockdown. *Archives of Disease in Childhood 106*, 8 (Aug. 2021), 791–797.

[4] Boonstra, T. W., Nicholas, J., Wong, Q. J., Shaw, F., Townsend, S., and Christensen, H. Using mobile phone sensor technology for mental health research: Integrated analysis to identify hidden challenges and potential solutions. *Journal of Medical Internet Research 20*, 7 (July 2018), e10131.

[5] Bry, L. J., Chou, T., Miguel, E., and Comer, J. S. Consumer smartphone apps marketed for child and adolescent anxiety: A systematic review and content analysis. *Behavior Therapy 49*, 2 (Mar. 2018), 249–261.

[6] Byrne, M. L., Lind, M. N., Horn, S. R., Mills, K. L., Nelson, B. W., Barnes, M. L., Slavich, G. M., and Allen, N. B. Using mobile sensing data to assess stress: Associations with perceived and lifetime stress, mental health, sleep, and inflammation. *DIGITAL HEALTH 7* (Jan. 2021), 20552076211037227.

[7] Caldeira, C., Chen, Y., Chan, L., Pham, V., Chen, Y., and Zheng, K. Mobile apps for mood tracking: an analysis of features and user reviews. *American Medical Informatics Association Symposium* (2018).

[8] Cao, J., Truong, A. L., Banu, S., Shah, A. A., Sabharwal, A., and Moukaddam, N. Tracking and predicting depressive symptoms of adolescents using smartphone-based self-reports, parental evaluations, and passive phone sensor data: Development and usability study. *JMIR Mental Health 7*, 1 (Jan. 2020), e14045. Company: JMIR Mental Health Distributor: JMIR Mental Health Institution: JMIR Mental Health Label: JMIR Mental Health publisher: JMIR Publications Inc., Toronto, Canada.

[9] Carlo, A. D., Hosseini Ghomi, R., Renn, B. N., and Areán, P. A. By the numbers: ratings and utilization of behavioral health mobile applications. *NPJ Digit Med 17* (2019), 54.

[10] Chikersal, P., Doryab, A., Tumminia, M., Villalba, D. K., Dutcher, J. M., Liu, X., Cohen, S., Creswell, K. G., Mankoff, J., Creswell, J. D., Goel, M., and Dey, A. K. Detecting depression and predicting its onset using longitudinal symptoms captured by passive sensing: A machine learning approach with robust feature selection. *ACM Transactions on Computer-Human Interaction 28*, 1 (Feb. 2021), 1–41.

[11] Dong, E., Du, H., and Gardner, L. An interactive web-based dashboard to track covid-19 in real time. *The Lancet Infectious Diseases 20*, 5 (May 2020), 533–534.

[12] Duncan, M., Murawski, B., Short, C. E., Rebar, A. L., Schoeppe, S., Alley, S., Vandelanotte, C., and Kirwan, M. Activity trackers implement different behavior change techniques for activity, sleep, and sedentary behaviors. *Interactive Journal of Medical Research 6*, 2 (Aug. 2017), e6685.

[13] Ezpeleta, L., Navarro, J. B., De La Osa, N., Trepat, E., and Penelo, E. Life conditions during covid-19 lockdown and mental health in spanish adolescents. *International Journal of Environmental Research and Public Health 17*, 19 (Oct. 2020), 7327.

[14] Ferreira, D., Kostakos, V., and Dey, A. K. Aware: mobile context instrumentation framework. *Frontiers in ICT 2* (2015), 6.

[15] Fitbit. Inspire hr 101 guide.

[16] Giménez-Dasí, M., Quintanilla, L., Lucas-Molina, B., and Sarmento-Henrique, R. Six weeks of confinement: Psychological effects on a sample of children in early childhood and primary education. *Frontiers in Psychology 11* (Oct. 2020), 590463.

[17] Giordan, L. B., Tong, H. L., Atherton, J. J., Ronto, R., Chau, J., Kaye, D., Shaw, T., Chow, C., and Laranjo, L. The use of mobile apps for heart failure self-management: Systematic review of experimental and qualitative studies. *JMIR Cardio 6*, 1 (March 2022), e33839.

[18] Grist, R., Porter, J., and Stallard, P. Mental health mobile apps for preadolescents and adolescents: A systematic review. *Journal of Medical Internet Research 19*, 5 (May 2017), e7332.

[19] Hamilton, J. L., Dreier, M. J., Caproni, B., Fedor, J., Durica, K. C., and Low, C. A. Improving the science of adolescent social media and mental health: Challenges and opportunities of smartphone-based mobile sensing and digital phenotyping. *Journal of Technology in Behavioral Science* (Oct. 2024).

[20] Hertz, M. F., and Barrios, L. C. Adolescent mental health, covid-19, and the value of school-community partnerships. *Injury Prevention 27*, 1 (Feb. 2021), 85–86.

[21] Hu, Y., and Qian, Y. Covid-19 and adolescent mental health in the united kingdom. *Journal of Adolescent Health 69*, 1 (July 2021), 26–32.

[22] Jones, E. A. K., Mitra, A. K., and Bhuiyan, A. R. Impact of covid-19 on mental health in adolescents: A systematic review. *International Journal of Environmental Research and Public Health 18* (2021).

[23] Jones, S. E., Ethier, K. A., Hertz, M., DeGue, S., Donna Le, V., Thornton, J., Lim, C., Dittus, P. J., and Geda, S. Mental health, suicidality, and connectedness among high school students during the covid-19 pandemic — adolescent behaviors and experiences survey, united states, january–june 2021. *MMWR Supplements 71* (2022).

[24] Larsen, L., Helland, M. S., and Holt, T. The impact of school closure and social isolation on children in vulnerable families during covid-19: a focus on children's reactions. *European Child Adolescent Psychiatry 31*, 8 (Aug. 2022), 1–11.

[25] Lecuelle, F., Leslie, W., Huguelet, S., Franco, P., and Putois, B. Did the covid-19 lockdown really have no impact on young children's sleep? *Journal of Clinical Sleep Medicine 16*, 12 (Dec. 2020), 2121–2121.

[26] Lehtimaki, S., Martic, J., Wahl, B., Foster, K. T., and Schwalbe, N. Evidence on digital mental health interventions for adolescents and young people: Systematic overview. *JMIR Mental Health 8*, 4 (Apr. 2021), e25847.

[27] Likhitweerawong, N., Boonchooduang, N., Kittisakmontri, K., Chonchaiya, W., and Louthrenoo, O. Effectiveness of mobile application on changing weight, healthy eating habits, and quality of life in children and adolescents with obesity: a randomized controlled trial. *BMC Pediatrics 21*, 1 (Nov. 2021), 499.

[28] Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., and Lee, S.-I. From local explanations to global understanding with explainable ai for trees. *Nature Machine Intelligence 2*, 1 (2020), 2522–5839.

[29] MacLeod, L., Suruliraj, B., Gall, D., Bessenyei, K., Hamm, S., Romkey, I., Bagnell, A., Mattheisen, M., Muthukumaraswamy, V., Orji, R., and Meier, S. A mobile sensing app to monitor youth mental health: Observational pilot study. *JMIR mHealth and uHealth 9*, 10 (Oct. 2021), e20638.

[30] Magson, N. R., Freeman, J. Y. A., Rapee, R. M., Richardson, C. E., Oar, E. L., and Fardouly, J. Risk and protective factors for prospective changes in adolescent mental health during the covid-19 pandemic. *Journal of Youth and Adolescence 50*, 1 (Jan. 2021), 44–57.

[31] Maharjan, S. M., Poudyal, A., van Heerden, A., Byanjankar, P., Thapa, A., Islam, C., Kohrt, B. A., and Hagaman, A. Passive sensing on mobile devices to improve mental health services with adolescent and young mothers in low-resource settings: the role of families in feasibility and acceptability. *BMC Medical Informatics and Decision Making 21*, 1 (Apr. 2021), 117.

[32] Mullick, T., Radovic, A., Shaaban, S., and Doryab, A. Predicting depression in adolescents using mobile and wearable sensors: Multimodal machine learning–based exploratory study. *JMIR Formative Research 6*, 6 (June 2022), e35807.

[33] Munos, B., Baker, P. C., Bot, B. M., Crouthamel, M., de Vries, G., Ferguson, I., Hixson, J. D., Malek, L. A., Mastrototaro, J. J., Misra, V., Ozcan, A., Sacks, L., and Wang, P. Mobile health: the power of wearables, sensors, and apps to transform clinical trials. *Annals of the New York Academy of Sciences 1375*, 1 (2016), 3–18.

[34] National Health Service. Anxiety in children, Feb. 2021.

[35] National Heart, Lung, and Blood Institute. Study reveals wearable device trends among u.s. adults, June 2023.

[36] Octavius, G. S., Silviani, F. R., Lesmandjaja, A., Angelina, and Juliansen, A. Impact of covid-19 on adolescents' mental health: a systematic review. *Middle East Current Psychiatry 27*, 1 (Dec. 2020), 72.

[37] Panchal, U., Salazar De Pablo, G., Franco, M., Moreno, C., Parellada, M., Arango, C., and Fusar-Poli, P. The impact of covid-19 lockdown on child and adolescent mental health: systematic review. *European Child Adolescent Psychiatry 32*, 7 (July 2023), 1151–1177.

[38] Pew Research Center. Fact sheets: Tech adoption trends, Nov. 2024.

[39] Pickel, G. 13 counties in southwest pa. will be next to reopen. *Penn Live* (May 2020).

[40] Satullo, S. K. Coronavirus in pennsylvania: 25% of state's 667 news cases in allegheny county as total creeps to 89k | coronavirus pa. map (073020). *lehighvalleylive.com* (July 2020).

[41] Seabold, S., and Perktold, J. statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference* (2010).

[42] Spadaro, B., Martin-Key, N. A., and Bahn, S. Building the digital mental health ecosystem: Opportunities and challenges for mobile health innovators. *Journal of Medical Internet Research 23*, 10 (Oct. 2021), e27507.

[43] Spitzer, R. L., Kroenke, K., Williams, J. B., and Löwe, B. A brief measure for assessing generalized anxiety disorder: the gad-7. *Archives of internal medicine 166*, 10 (2006), 1092–1097.

[44] Staff, N. Pennsylvania, delaware close all schools due to outbreak. *NBC10 Philadelphia* (March 2020).

[45] Staner, L. Sleep and anxiety disorders. *Dialogues in clinical neuroscience 5*, 3 (2003), 249–258.

[46] Vega, J., Li, M., Aguillera, K., Goel, N., Joshi, E., Khandekar, K., Durica, K. C., Kunta, A. R., and Low, C. A. Reproducible analysis pipeline for data streams: Open-source software to process data collected with mobile devices. *Frontiers in Digital Health 3* (Nov. 2021).

[47] World Health Organization. Anxiety disorders.

[48] Wu, M., and Luo, J. Wearable technology applications in healthcare: A literature review. *Online Journal of Nursing Informatics* (Nov. 2019).

[49] Xu, X., Yao, B., Dong, Y., Gabriel, S., Yu, H., Hendler, J., Ghassemi, M., Dey, A. K., and Wang, D. Mental-llm: Leveraging large language models for mental health prediction via online text data. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 8*, 1 (Mar. 2024).

Received 6 January 2025; Accepted 18 February 2025