Utilizing Passive Data Collection to Detect Anxiety and Depression

ALDRICK JOHAN, University of Virginia
KESHAV AILANEY, University of Virginia
JOHAN KETKAR, University of Virginia
WEI WANG, University of Virginia

Using passive data collected from smartphones, daily behavior is modeled through images of two-dimensional data. Such is applied to k-means clustering in order to illustrate significant differences in behavior and through the use of a convolutional neural network, the images were utilized to predict both anxiety and depression among users based on PHQ and GAD scores. The experiment yielded 95% and 92% accuracy, respectively.

CCS Concepts: • **Deep Learning** → *Convolutional Neural Networks*; • **Clustering** → *Anomaly Detection.*

Additional Key Words and Phrases: anxiety, depression, neural networks, clustering, smart phones

1 INTRODUCTION

Early detection of major depressive disorder or depression in an individual can lead to more effective treatment plans and significantly increase the probability of recovery. However, the diagnose and detection of depression is based on physical examination, evaluation, or lab testing results. These methods may require both time and financial resources from the patient, thus, contributing to the overall difficulties of the diagnosing process. The goal of this technical project is to apply behavior modeling techniques to detect depression and anxiety from passive data collected from users' smartphones. Specifically, machine learning algorithms like clustering and classification will be employed to complete such a task. These methods will be used to distinguish the differences in users' behavioral data. This approach can greatly reduce the amount of time commitment required from the users and simplify the process of detecting depression.

2 RELATED WORK

Machine learning algorithms have been applied in many behavioral modeling studies. For example, Srividya et al. (2017) applied several classification and clustering techniques to model behavior for mental health. The dataset used in this study was unlabeled and was first analyzed with clustering algorithms to generate group labels. The clustering algorithms used in this process included K-means, hierarchical, and K-medoids. The generated group labels were then validated by computing the Mean Opinion Score. With the labeled data, classification algorithms were performed to build a prediction model for mental health. The algorithms involved in this process included logistic regression, Naïve Bayes, support vector machine (SVM), decision tree, KNN, ensemble (bagging), and random forest tree. The results indicated that KNN, SVM, and random forest trees had achieved similar performance. The accuracy score of all three

Manuscript submitted to ACM

models was about 90%. However, this project had not considered utilizing a neural network model for clustering or classification.

Altum et. al (2010) models and classifies physical human behavior using body sensors and a variety of classification techniques, respectively. Altum et. al (2010) attempts to utilize such sensors to provide a more convenient means of activity detection rather than commonly implemented visual based systems which require adequate lighting and equipment among other factors. Activities such as standing, sitting, walking, and specific forms of exercise were classified using bayesian, decision trees, least squares method, k nearest neighbors, dynamic time warping, support vector machines, and artificial neural networks. Feature extraction, reduction, and cross validation were also implemented. The most effective model was Bayesian, but many yielded high accuracies. For example, the ANN yielded 96.2% accuracy. However, Album et. al (2010) did not consider the usage of a CNN to model the behavior of users and sensor data, as exhibited in the following paper.

Chikersal et al. (2021), utilizing the same AWARE framework as our research, constructed a feature extraction technique with a prediction accuracy of 85% for depression. This study involved several different machine learning algorithms for a variety of purposes. For example, the dataset used in this project contained several feature sets, including call and location features, and randomized logistic regression was utilized to select the feature sets most relevant for training the models. logistic regression was also used along with Gradient Boosting Classifier in the model training and validation process. Lastly, an ensemble classifier, using AdaBoost with Gradient Boosting Classifier as the base estimator, was utilized for depression detection. Classification techniques were also employed in our study; however, a CNN model was utilized for such a task.

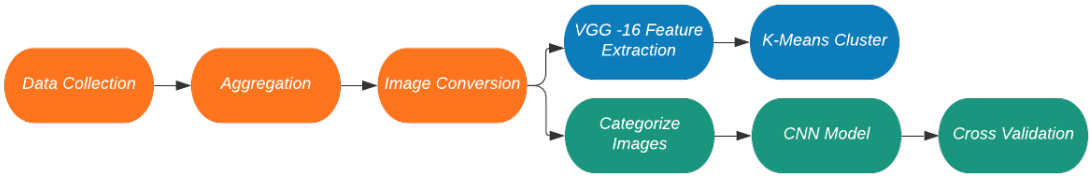

Fig. 1. Process overview

3 METHODS

This experiment was consisted of two main phases: data processing and machine learning. In the data processing phase, data was categorized, aggregated, and converted into images. Then, the images were labeled for supervised learning. In the machine learning phases, clustering and CNN models were constructed and trained with the images. An overview of the process is shown in figure 1.

3.1 Data Collection

Using AWARE [4], a framework to log user mobile activity for research, valuable data were collected by participants at Christopher Newport University in 2020 and 2021. Information relating to the user's calls, conversations, activities, location, Wi-Fi connection, and screen time were gathered by utilizing sensors and plug-ins available on the participants' devices [4] to track their behavior. The dataset contained mobile contextual information collected from 48 unique iOS or Google Android devices. For each sensor/plug-in type, there were certain unique features. For example, location

data contained the 'double_longitude' feature and the call data contained the 'call_duration' feature. However, there were also some shared features in all types of data. These common features included 'name', 'timestamp', 'device_id', and 'device_label'. The 'name' feature identified the data type. Some example values for this feature were "calls" and "locations" for call data and location data, respectively. The 'timestamp' feature contained the number of milliseconds between 1970 and the time for when the data entry was recorded. The 'device_id' was a unique identifier assigned to each device while the 'device_lable' was an identifier assigned to each participant. Therefore, if a participant had two devices, data collected from these two devices would have different 'device_id' values but the same 'device_lable'.

As a means to understand their mental health, each user was asked to complete weekly PHQ-9 and GAD-7 surveys, which provided ground truth to determine their degree of depression and anxiety, respectively. PHQ-9 [5] is a self-administered survey that is consisted of 9 individual questions, and each question requires the respondent to provide a score between 0 and 3. In total, the PHQ-9 score ranges from 0 to 27, and the score is interpreted as follows: no depression (0 – 4), mild depression (5 – 9), moderate depression (10 – 14), moderately severe depression (15 – 19), and severe depression (20 – 27). The effectiveness of the PHQ-9 survey has been validated in two large studies, and it is half the length of many other evaluation methods [5]; thus, it is used as the ground truth for depression measures in this study. Similarly, the GAD-7 [6] is a self-report questionnaire for anxiety measures that is consisted of 7 items. For each item, the respondent would need to provide a score between 0 and 3, corresponding to 'not at all', 'several days', 'more than half the days', and 'nearly every day', respectively. The total score ranges from 0 to 21, and the interpretations for the score are as follows: no anxiety (0 – 4), mild anxiety (5 – 9), moderate anxiety (10 – 14), and severe anxiety (15 – 21).

3.2 Data Aggregation

The data was organized according to the user labels and sensor/plug-in types, such as user label id "203BN" and plug-in type iOS activity. Next, the data was further separated based on the date it were collected and then, aggregated into hourly intervals. The aggregation methods used in this process included finding the sum, mean, mode, or number of data features that occurred during each hourly interval.

3.3 Feature Extraction

Additionally, certain new features were extracted. For example, the numbers of incoming calls, connected calls, and disconnected calls were computed based on existing features of the call data.

The methods of aggregation employed and any data extractions performed for each type of data are listed below:

- Call data: extract these new features from "call_type": "num_incoming", "num_connected", "num_dialing", "num_disconnected", "num_calls". Each of the new features describes the number of incoming calls, connected calls, dialing/outgoing calls, and disconnected calls. Find the sum of "call_duration". Find the mode for "call_type" and "trace". For all other features, including the extracted ones, count the number of occurrences.
- Location data: find the mode for "provider". For all other features, use mean for aggregation.
- Audio data: compute the length of each conversation ("convo_length") using the "double_convo_start" and "double_convo_end" features. Find the mode for the following features: "datatype", "inference". Find the mean value for these features: "double_convo_end", "double_convo_start", "double_energy". Count the number of conversations occurred ("num_convos") based on "double_convo_start".

- Activity data: based on “activity_name”, count the total number of occurrences for each of the following activity types: “cycling”, “stationary”, “running”, “walking”, “automotive”, “unknown”. Compute the mean value for the “confidence” feature.
- Screen data: count the number of times the device is picked up by the user (“num_pickups”) based on “screen_status”. Find the total amount of screen time based on “screen_status”.
- Wi-Fi data: count the number of unique “bssid”. Find the mode for “security” and “ssid”. Find the mean value for “frequency” and “ssid”
- Sensor Wi-Fi data: count the number of unique “bssid”. Find the mode for “mac_address” and “ssid”.

3.4 Image Generation

Another step of the data processing was to convert the data into images. Once the raw data were separated into dates and aggregated into hourly intervals, each data value was scaled to be between 0 and 255, representing pixel values. Each data table contained 24 rows, corresponding to the 24 hours of a day, and the column dimension varied based on the data features available for each sensor/plug-in type. The dimension of the images would correspond with the dimension of the data tables. However, since the column dimension differed for each data type, images were resized to be 32 x 32.

Figure 2 contains an example images generated using the activity data collected from user 25BY, which is shown in table 1. The figure contains some color spots that are noticeably “lighter” compared to others. The color difference was caused by the original data values. During the image generation process, the data value was directly converted into pixel values. In this case, a higher original value would create a deeper color pixel. Additionally, it was apparent that most of the color spots formed columns. These columns corresponded to the features contained in the original data set. As shown in table 1, the "stationary" feature contained the highest data values, which described the number of times the user was being still during a day; thus, the most clear column in figure 2 corresponded to the "stationary" feature. From the color difference, it was possible to determine which activity was more prevalent that other.

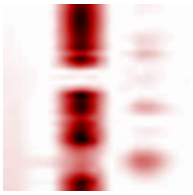

Fig. 2. Activity data image from user 25BY

For supervised learning algorithms, the data should be labeled. In our study, the results of the weekly PHQ-9 and GAD-7 surveys were used to label the data. The PHQ-9 scores ranged from 0 – 27, and the scores were interpreted according to this guideline: minimal depression (0 – 4), mild depression (5 – 9), moderate depression (10 – 14), moderately severe depression (15 – 19), and severe depression (20 – 27). The GAD-7 scores ranged from 0 – 21 and were interpreted as follows: minimal anxiety (0 – 4), mild anxiety (5 – 9), moderate anxiety (10 – 14), and severe anxiety (15 – 21). Since the surveys were conducted on a weekly basis, all data collected between the previous survey and the current survey would be labeled using results from the current survey. For example, if a survey was submitted on 10/10/2020 and the

cycling	stationary	running	unknown	walking	automotive
0	224	0	0	12	0
0	244	0	0	0	0
0	236	0	0	2	0
.	.	.	.	.	.
.	.	.	.	.	.
0	222	0	0	24	0
0	222	0	0	0	0

Table 1. Aggregated activity data from user 25BY

previous survey was conducted on 10/03/2020, data that occurred between 10/10/2020 and 10/04/2020 would be labeled with the results from the 10/10/2020 survey.

3.5 Behavioral Image Clustering

In this study, both unsupervised and supervised learning methods were employed to model users' behavior and detect signs of depression and anxiety. The unsupervised learning algorithm used was clustering, specifically K-Means, and the supervised learning algorithm used was the Convolutional Neural Network (CNN).

3.5.1 Clustering of Individual Behavior. K-Means clustering, an algorithm to partition data into k clusters based on similarity, and a pre-trained model were used to analyze behavior patterns per user per sensor and amongst all users per sensor.

The pre-trained model was a VGG16 convolutional neural network. The model was used to extract a feature vector from each image and then the K-Means algorithm clustered images based on how similar the feature vectors are.

In each instance of clustering, the optimal number of clusters was determined by using silhouette score analysis. Silhouette score analysis is calculating the mean intra-cluster distance a and the mean nearest-cluster distance b for each image after clustering with a specific k value. A silhouette score is determined by taking the average of $\frac{b-a}{max(a,b)}$ for all samples. Silhouette scores were determined for k values 2 through 9 inclusive. Only the clustering output with the parameter value k which corresponded to the highest silhouette score was reported in the results.

Images were sorted into buckets where each bucket only contained images that corresponded to one user and one sensor type. For example, all images in one bucket belonged to the user with device label '25BY' and were generated from the 'audio' sensor. The images in each bucket were then clustered using the K-Means algorithm and the results corresponding to the highest silhouette score were reported. The results of this clustering were used to analyze the behavior patterns of each user individually per each sensor.

3.5.2 Clustering of Group Behavior. Images were also sorted into buckets where each bucket only contained images that corresponded to one sensor. For example, all images in one bucket were generated from the 'audio' sensor but could have belonged to any user. The images in each bucket were then clustered using the K-Means algorithm and the results corresponding to the highest silhouette score were reported. The results of this clustering were used to analyze behavior patterns of all users per each sensor.

3.6 Behavioral Image Classification

A convolutional neural network, typically used for image classification, was also utilized to analyze the correlation between user activity and mental health. Specifically, it was used to determine the optimal set of input sensors for predicting a user's mental health. The CNN used for this research was made of the following layers: a batch normalization layer, a layer normalization layer, multiple max pooling layers, multiple 2D convolutional layers, a dropout layer, a flattening layer, a 'relu' activation layer, and a 'softmax' activation layer. The model was compiled using the 'nadam' optimizer and used sparse categorical cross entropy to calculate loss. For use in this model, the data was labeled using the user's depression or anxiety classification for that week. The data was in the form of 32 x 32 images, and each image represented one day of a user's activities collected from one sensor. Two testing methods were used to determine which sensors provided the best information for predicting the depression or anxiety levels of a user: a regular test using a 80%-20% train-test split of the data and then another test using leave-one-out cross-validation. For both of these tests, the CNN attempted to classify test data based on the data that was used to train it.

To begin, the 80%-20% split was tested for all sensor combinations. The model was trained using 80% of the available data and was tested using 20% of the available data. The available data for each test was the entirety of the data provided by each of the input sensors. Each combination of sensors was run for 20 epochs. Each sensor combination was tested with two sets of categories: the anxiety categories provided by the gad7 survey and the depression categories provided by the phq9 survey. The final accuracy of the model was used to determine how effective the combination of sensors was at predicting a user's mental health in regards to anxiety or depression. The accuracy for these tests were low due to the relatively small data set.

Consequently, leave-one-out cross-validation was used for testing. For these tests, there were some adjustments made to the method. First, the split of testing and training data was changed to use the leave-one-out method. This involved excluding one user's data from the training data, and using it for the test data. The accuracy for this test is noted and then the process is repeated with another user. Once all the users have been excluded once, the average accuracy from all the tests is calculated. Another change was with the combinations of input sensors that were used. For the cross-validation tests, a maximum of three sensors were used as an input. There were two reasons for this decision. First, during the testing of the 80%-20% split it was found that using four or more sensors degraded the performance of the model sharply. Second, using leave-one-out cross-validation was very time consuming due to the large amount of tests that had to be run for each combination of sensors. Furthermore, some sensors had separate versions for iOS and Android, such as 'ios_activity' and 'google_activity.' These sensors were considered as one sensor for the purposes of this test. The final change made to the method for cross-validation was reducing the number of epochs from 20 to 10. This change was made to reduce runtime and because there was not a large increase in accuracy between the 10th and 20th epoch.

4 RESULTS

4.1 Behavioral Image Clustering

4.1.1 Individual Clustering. All images belonging to a single user generated from a single sensor were clustered according to the methods above. Table 2 shows the average number of clusters reported.

Table 2. Average Clusters

Sensor	# of Clusters
screen	1.92
ios_activity	2.05
google_activity	1.88
audio	2.26
audio_android	2.50
locations	3.44
calls	2.65

An example of the produced clustering is illustrated in figure 3, depicting the differences in a user's activity. For images in the first cluster, the high intensity of values found solely in the middle column represent the sedentary behavior of the user while the second cluster illustrates records of walking and automotive behavior as found in the last columns, respectively.

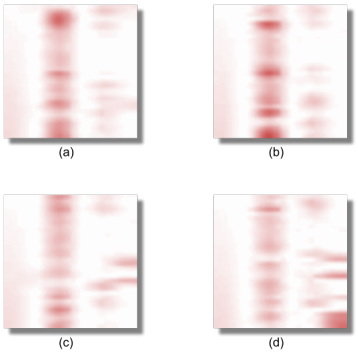

Fig. 3. User 118LZ Activity Clustering

4.1.2 Group Clustering. All images from every user belonging to the same sensor were clustered according to the methods above. This was repeated for every sensor and the number of clusters with the highest silhouette score are reported in table 3.

Table 3

Sensor	# of Clusters	Silhouette Score
screen	2	0.2908
ios_activity	2	0.1235
audio	2	0.5191
locations	2	0.4196
calls	2	0.5389
audio_android	2	0.4624
google_activity	2	0.3221

An example of such clustering is illustrated below in Figure 4. Although difficult to interpret, the images found in the second row represent the second of two clusters. Such images contain more recordings of sedentary behavior as opposed to the first cluster and thus, were the basis of the clustering.

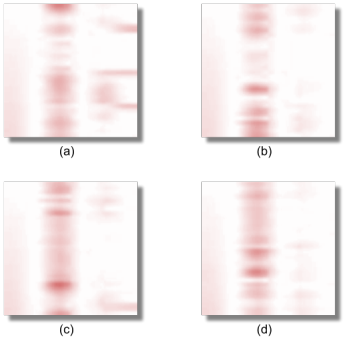

Fig. 4. Sample Clustered Images of Activity Sensor

When analyzing the results of this clustering it was observed that some users had a majority of images from a sensor belonging to the same cluster. This indicates that such a users behavior, as measured by the sensor was consistent between days. Other users had images evenly split between clusters, suggesting that their behavior was inconsistent between days.

Table 4 shows the portion of images belonging to a single cluster for two different users. User 49ZR's images were more likely to belong to the same cluster, meaning user 49ZR's behavior was more consistent across different days. User 25BY's images were more likely to be split evenly between clusters, meaning user 25BY's behavior was less consistent across different days.

Table 4. Portion of images belonging to dominant cluster

Sensor	49ZR	25BY
calls	.90	.61
locations	.99	.58
audio	.78	.71
ios_activity	.94	.56
screen	.78	.55

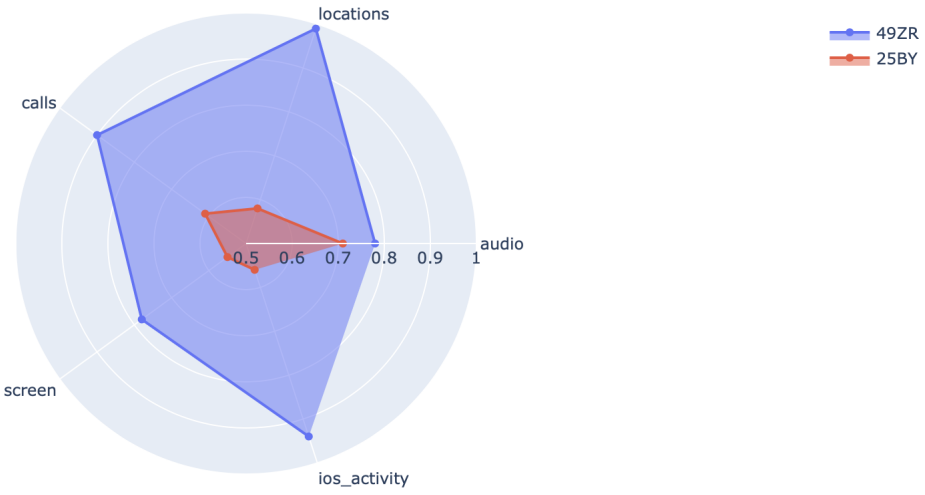

Fig. 5. Portion of images belonging to dominant cluster

4.2 Behavioral Image Classification

4.2.1 80%-20% Split.

The results in table 5 and table 6 were acquired while using a 80%-20% split of train and test data, to calculate the accuracy of the CNN. The 5 input combinations with the highest accuracy are shown in the table below in descending order. The results for both the gad7 categories and the phq9 categories are shown.

Table 5. GAD Results

Rank	Inputs	Accuracy
1	google_activity	0.5652
2	audio_android, google_activity	0.5124
3	audio_android	0.4712
4	audio, google_activity	0.4702
5	ios_activity, google_activity, audio_android	0.4439

Table 6. PHQ Results

Rank	Inputs	Accuracy
1	google_activity	0.6232
2	audio_android, google_activity	0.5992
3	audio_android	0.5481
4	audio, audio_android, google_activity	0.5193
5	ios_activity, google_activity, audio_android	0.4849

4.2.2 *Cross-Validation Results.*

The results contained in table 7 and table 8 were acquired while using cross-validation to calculate the accuracy of the CNN. The 5 input combinations with the highest accuracies are shown in the table below in descending order. The results for both the gad7 categories and the phq9 categories are shown. For these results, note that both ios_activity and google_activity are considered the same sensor. This also applies to audio_android and audio.

Table 7. GAD Results

Rank	Inputs	Accuracy
1	ios_activity, google_activity	0.9517
2	ios_activity, google_activity, screen	0.8820
3	screen	0.8809
4	screen, audio_android, audio	0.8348
5	ios_activity, google_activity, screen, audio_android, audio	0.8258

Table 8. PHQ Results

Rank	Inputs	Accuracy
1	ios_activity, google_activity	0.9287
2	screen	0.9074
3	ios_activity, google_activity, audio_android, audio	0.8685
4	ios_activity, google_activity, screen	0.8677
5	ios_activity, google_activity, screen, audio_android, audio	0.8528

5 CONCLUSION

Using a combination of VGG16 feature extraction and K-Means clustering on sensor data reveals differences in user's behavior patterns as well as the ability to recognize a user's tendencies. The silhouette scores in table 3 demonstrate which sensor's reveal the most significant behavioral differences in this data. For example, images in different clusters of the call data were more different than images in different clusters of the ios_activity data. Further analysis of a user's distribution of images across clusters is an effective method to determining if the user acts consistently or inconsistently across different days.

The results produced by the CNN show that it is a good predictor of anxiety and depression if these two conditions are met: enough data is provided to the model and the correct sensors are provided as an input. The CNN tended to provide more accurate predictions when more data was provided. This by comparing the results in table 5 with the results in table 7. The results in table 5 utilize a 80%-20% split of training and test data while the results in table 7 uses the leave-one-out method. The leave-one-out method provides more data as it includes every user's data except for one, demonstrating the relationship between amount of data and accuracy. The second condition for accurate predictions is that the correct sensors have to be used as the input. This conclusion can be deduced by looking at any of the prior CNN results tables. Activity and audio data is found near the top of all the results in terms of accuracy. When these two conditions are fulfilled, a CNN can be used as a good predictor of the anxiety or depression levels of people.

As a result, modeling behavior data as images can be applied to clustering methods or CNNs for analysis or prediction, respectively.

6 REFERENCES

[1] Srividya, M., Mohanavalli, S. and Bhalaji, N., 2018. Behavioral Modeling for Mental Health using Machine Learning Algorithms. *Journal of Medical Systems*, 42(5).

[2] Chikersal, P., Doryab, A., Tumminia, M., Villalba, D., Dutcher, J., Liu, X., Cohen, S., Creswell, K., Mankoff, J., Creswell, J., Goel, M. and Dey, A., 2021. Detecting Depression and Predicting its Onset Using Longitudinal Symptoms Captured by Passive Sensing. *ACM Transactions on Computer-Human Interaction*, 28(1), pp.1-41.

[3] Altun, K., Barshan, B. and Tunçel, O., 2021. *Comparative study on classifying human activities with miniature inertial and magnetic sensors.*

[4] Ferreira, D., Kostakos, V. and Dey, A., 2015. AWARE: Mobile Context Instrumentation Framework. *Frontiers in ICT*, 2.

[5] Kroenke, K. and Spitzer, R., 2002. The PHQ-9: A New Depression Diagnostic and Severity Measure. *Psychiatric Annals*, 32(9), pp.509-515.

[6] Williams, N., 2014. The GAD-7 questionnaire. *Occupational Medicine*, 64(3), pp.224-224.