Spring 5-10-2019

DepressionGNN: Depression Prediction using Graph Neural Network on Smartphone and Wearable Sensors

Param Bidja
param.bidja@uconn.edu

DepressionGNN: Depression Prediction using Graph Neural Network on Smartphone and Wearable Sensors

Honors Scholar Thesis
Author: Param Bidja
Major Advisor: Dr. Jinbo Bi
Associate Advisor: Chao Shang

Department of Computer Science and Engineering, University of Connecticut, Storrs, USA
param.bidja@uconn.edu, jinbo.bi@uconn.edu, chao.shang@uconn.edu

Abstract. Depression prediction is a complicated classification problem because depression diagnosis involves many different social, physical, and mental signals. Traditional classification algorithms can only reach an accuracy of no more than 70% given the complexities of depression. However, a novel approach using Graph Neural Networks (GNN) can be used to reach over 80% accuracy, if a graph can represent the depression data set to capture differentiating features. Building such a graph requires 1) the definition of node features, which must be highly correlated with depression, and 2) the definition for edge metrics, which must also be highly correlated with depression. In this analysis, we find those node features and edge metrics which lead to graph that accurately represents a depression data set on which GNN can achieve over 80% accuracy.

1 Introduction

1.1 Motivation of Study

The motivation of this study is to determine if a novel machine learning technique can accurately detect depression signals based on physical and GPS sensing data. The problem boils down to a node classification problem, for which a new technique has proven to be extremely accurate: Graph Neural Networks (GNN). We want to explore the usefulness of Graph Neural Networks on depression prediction because depression is not an easily classifiable disease. Even in its current state, depression prediction is not clear-cut and easily understood, because depression has so many factors. It's a disease that impacts its patients in physical, mental, and social ways, and a disease for which the symptoms can manifest in many different ways [3]. There is no ”single-formula” to spot depression, which makes it a fairly complicated classification problem - too complex for traditional classification approaches. Thus, the motivation of this study is to utilize the novel classification technique of Graph Neural Networks on the problem of depression prediction.

1.2 Depression

Depression is one of the most prevalent mental health disorders in the world. In the U.S. alone, approximately 17.3 million adults, more than 7% of the entire adult population, suffers from major depression disorder [2]. Those who suffer from depression come from many diverse backgrounds, as it can impact any demographic. Beyond the impact on an international level, the disease is extremely expensive for individuals and requires persistence by skilled clinicians to solve. Skilled clinicians cannot always be available, especially in developing parts of the world [4]. Most countries lack adequately trained health professionals.

Health professionals use a set of signals that indicate depression levels, which can often range from physical features, to behavioral, to social [6]. In this study we try to capture those features for depression level prediction. We attempt to classify individuals' depression levels based on their physical activity, sleep, heart rate, and GPS data. This sensing data gives us insight into the features that define depression candidates most. We will show how a technique called Graph Neural Networks thrives in such a classification problem, especially when the problem domain is well-represented by a graph (or multiple graphs), so much so that we will get above 80% test accuracy in our data set.

1.3 Hypothesis

Before we dove into the methods and analysis, we developed hypotheses based on our intuition and understanding of depression. These hypotheses included:

1. Depression is correlated with physical features, namely activity metrics, sleep metrics, and heart rate metrics.
2. Patients of depression in a college environment exhibit similar spatial patterns of movement, meaning the types of places they visit (or do not visit) are similar.
3. Graph Neural Networks can accurately classify depression candidates based on physical features and spatial trajectories

We made hypothesis 1 because depression manifests in physical signals, because doctors look for specific signals when diagnosing depression. Those physical signals include, but are not limited to, activity, sleep, and heart rate information. Also, self-identification surveys are commonly used which use physical signals as a means of depression level classification.

Hypothesis 2 says that depression patients in a college environment, which our study's patients are confined to, will exhibit similar spatial patterns. This intuitively makes sense, because depression is largely a social disease, and where a person goes says a lot about their social behavior. For example, those who visit classrooms are more likely to interact with others compared to those who stay in their residence areas. Also, a college student's movement gives insight into their physical activity levels, as students who visit the gym, sports fields, etc. are more likely to be more active. Given the social and physical attributes of depression, and the correlation between spatial trajectories and social and physical attributes - we hypothesized that spatial features would be relevant.

Hypothesis 3 is rooted in the recent rise in graph neural networks. A plethora of literature has shown the power of graph neural networks, especially in the problem of node classification. Due to the success of the technique, and how well it suits the data set we've collected, we suspected it would be a good technique for node classification.

These hypotheses were rooted in intuition, however we use data and statistics to determine whether they are reasonable or not. In the end we confirmed some of these hypotheses to be true, while others to be false.

1.4 Terminology

As we move into the technical methods and results of this study, we will use some terminology repeatedly so we should clarify definitions for reference.

- **clinical result**: in our study, we gathered clinical result for each participant. These results are given by practicing medical doctors, and act as ”true labels” in our study. Doctors classify levels of depression based on a plethora of signals, of which we'd like our neural network to recognize the most important ones.
- **QUIDs**: the Quick Inventory of Depressive Symptomatology survey (QUIDS) is a survey for self-identifying depression levels. This is a survey that many depression patients take prior to receiving a clinical diagnosis, as a first step in the diagnosis [1].

- **trajectory**: we define trajectory as a sequence of geo-spatial information. This captures where an individual has traveled throughout their day.
- **categorical data**: categorical data means geo-spatial data that captures the types of buildings an individual visited. For example, categories like residential, entertainment, academic, etc. capture the types of buildings a college student visits.
- **edge distances**: not to be confused with Euclidean distance, we refer to edge distance as the distance between any two vertices in a graph. We will explore different metrics for distance, so do not assume distance to mean simple Euclidean distance.

2 Methods

2.1 Data Collection

The study gathered data from a group of 79 participants through 2 main sensors: Fitbit and smartphones. The data was collected from October 2015 to March 2017, however not continuously. The collected data can be broken into the following categories: QUIDS survey results, clinical diagnosis results, Fitbit sensing data, raw GPS data, categorical GPS data, and some auxiliary metadata such as wifi access point usage. The data was fully anonymized to preserve the privacy of the participants by assigning each participant a random ID. In the following, we first describe the various types of data in detail and then briefly describe the participants information

2.1.1 QUIDS Survey Results The Quick Inventory of Depressive Symptomatology (QUIDS) report is a self-driven report that helps an individual recognize their own, self-diagnosed, levels of depression. It consists of 30 questions that range from topics like activity, to sleep, to diet, and even emotional factors. It's an establish, commonly used metric in many depression classifications. The survey has a weighted scoring system that tells the patient, based on their survey responses, their probable level of depression.

This survey was taken by the participants in the study every week, so we have their self-diagnosis throughout the study. To ensure standardization amongst different participants, each participant also has a clinical, medical doctor given label for depression levels.

2.1.2 Clinical Data Before the study began, each participant received a physical diagnosis from a medical doctor regarding their depression level. The doctors classified each patient based on a plethora of attributes, including their physical symptoms, social patterns, QUIDS results, and many others. The participants received one clinical label at the start of their data collection, which was assumed to stay the same throughout the study. This means we assume a depressed individual stayed depressed throughout the study, and same for non-depressed. We will treat these clinical labels as ”true labels” during the supervised and unsupervised learning portions of this project.

The means of clinical diagnosis is an interview that was designed based on the Diagnostic and Statistical Manual of Mental Health (DSM-5) and QUIDS evaluation, the clinician classified individuals as either depressed or non-depressed during the initial screening. In addition, depressed participants had follow-up meetings with the clinician periodically (once or twice a month determined by the clinician) to confirm their self-reported QUIDS scores with their verbal report during the meetings.

2.1.3 Fitbit sensing data Fitbit sensors have capabilities to track three major categories of data: activity, sleep, and heart rate. In our study, activity and sleep were of the most interest, thus we analyzed two categories of data. The Fitbit data was collected by Fitbit, then sent to us in a formatted, reduced way so we did not get all raw data.

The record frequency of Fitbit data is regulated by the sensor itself, which was on average 1 activity record per day per participant, and 1 sleep record per day per participant. We assume high accuracy on these results, because Fitbit claims for activity and sleep data Fitbit sensors are ”highly accurate”.

2.1.4 Raw GPS Data Using smartphone GPS sensors and wifi location data, we gathered the GPS location data for each participant. All participants spent most of their time on the University of Connecticut campus (or surrounding areas), so we were able to gather GPS data based on wifi for a subset of the participants on campus.

The frequency of this data was dependent on sensor and network factors, like the operating system of the smartphone, or the frequency of polling by access points. On average, we received 24 points per day, or 2 per hour from each participant.

The raw GPS data gives us GPS coordinates with timestamps, along with some other metadata depending on how the data is collected (through wifi or cellular data). The wifi GPS points come with metadata regarding the specific access point on UConn’s campus that was used, which will be useful for the next section of data (categorical).

2.1.5 Categorical GPS Data The subset of the raw GPS data that was recorded while the participant was connected to a campus wifi access point (AP) has metadata about said AP. Using our university’s mapping of AP IP addresses to buildings, we were able to label these data points with specific buildings across our campus.

We manually classified each building into 5 distinct categories that summarize the types of buildings that exist on our campus:

1. Entertainment - buildings that are used for entertainment and social events, such as Jorgenson Theatre and Student Union
2. Activity - buildings that are used for physical activity and sports, such as the Student Recreation Facility and athletic fields
3. Academic - buildings where classes are held
4. Residential - on-campus residential halls
5. Library - buildings dedicated for studying, of which there are a few libraries. This category is unique because it’s where many students spend time throughout the day, between academic buildings.

Once we had a mapping of wifi GPS points to buildings, and buildings to categories, we could classify each GPS point collected over wifi to a building category. This adds a dimension of purpose to the data, giving us a distinction of why participants would be in certain areas.

2.1.6 Participants We recruited 25 Android users and 54 iPhone users, all students of UConn, aged 18-25. Among the 25 Android users, 6 were classified as depressed and 19 were classified as non-depressed. The Android phones were from a variety of manufacturers, including Samsung, Nexus, HTC, Xiaomi, Motorola and Huawei. Among the 54 iPhone users, 13 were classified as depressed and 41 were classified as non-depressed.

All participants used their own smartphones except for two participants (who did not have smartphones and borrowed Android phones from us). Of that data, we had errors on collecting 27% where those were empty or missing data due to errors in the collection process.

2.2 Visual Analysis using Heatmaps

The initial method we deployed for testing if GPS spatial trajectories had an impact on depression was a visual analysis through heatmaps. We divided the data into two subsets, one for the depressed patients and the other for the non-depressed patients. After normalizing the data using random selection to even out the size of the participant pool in the two sets, we created visual heatmaps to analyze spatial movement of the two groups over our campus. We developed such maps with a dimension of time, to see the daily movement for each group.

Such a method was useful for initial analysis because it gave insight into the different parts of campus that are routinely being visited by each group, which helped devise the category list for categorical GPS data. Said heatmaps are shown in the Results section.

2.3 Graph Neural Network

Graph Neural Networks are a generalization of convolution neural network (CNN) which are applied on graph structures [5]. GNNs have been recently used to handle many complicated tasks, like matrix completion, manifold analysis, and predictions on connectivity in social networks [7]. GNNs have the ability to recognize neighborhoods or pockets in the graph that have similar features, and classify the nodes in that neighborhood accordingly [**?**]. GNNs have proven to be very successful in node classification problems.

The performance of GNNs are directly correlated with the structure of the graph, because the graph learns its neural network weights based on the neighborhoods formed in the graph. Thus, forming a well-devised and ground truth graph is the key to good classification performance for GNNs.

Due to the complicated nature of depression, we will attempt to use a GNN on the data we've gathered. The difficult part will not be actually using the GNN, because a plethora of work has already been done in implementing GNNs, but rather the difficulty arises in building the graph. The remainder of this paper will discuss how the graph itself should be constructed.

2.4 Distance Metrics

To approach any problem with a Graph Neural Network method, the most important consideration that must be made involves how to construct the graphs. Constructing the graphs requires a metric to define the distance between nodes, and in our case the nodes represent the depression study participants, we need a metric to define the distance between participants.

There are countless ways to define the distance or similarity (which is the inverse of distance) between two participants, however we focused on the means that made the most objective sense. The main criteria for the best metrics is that it must maximize the distance between participants in the depressed and non-depressed groups while minimizing the distance between members of the same group. By fulfilling these two criteria, a metric will aid the GNN in it's pursuit to find definitive attributes that separate the two groups.

2.4.1 Data for Edges We analysed our data-set to understand which data markers most significantly separate the depressed and non-depressed subsets of participants. We chose not to consider Fitbit features as a feature for defining edges because those attributes were captured in each node. By including those Fitbit attributes in the edge connections, we would be redundantly using data throughout the graph or we would need to split that data used at the nodes, both of which would lower the performance of a GNN. Thus, we were left with one other part of the data to use: GPS data.

The GPS data can be broken down into two sets - the categorical GPS data and the raw GPS data. The categorical data is just a byproduct of processing done on the raw GPS data, however it gives us more information because it gives us the building category that each GPS point maps to. Thus, we considered the GPS categorical data as the data-set of choice for determining edge connections in our graph.

Now that we've focused on the categorical GPS data for edge distances, let's define how we will use that data when determining distances between participants. There are two ways we utilize the categorical data: a **time-series trajectory** representation and a **frequency representation**.

The time-series trajectory representation is a sequence of categories (the categories 1-5 defined above) based on where a participant traveled across a single day. Thus, a categorical trajectory is define per day, and contains integers in the range [1,5] based on the categories we defined. For example, a categorical trajectory of [1,3,2] represents a participant's travel for a single day across buildings, first visiting a building with category 1, then a building with category 3, then a building with category 2.

The frequency representation of a participant's categorical GPS data is a frequency table of their categorical travel for a single day. A daily vector V is created for each patient P where the ith element in V represents how many times patient P visited a building with category i. Recall that the categories are labeled 1-5, so these vectors all have a length of 5. For example, a frequency vector of 2, 1, 0, 3, 0 represents a participant's travel over a single day where he/she visited a building of category 1 twice, a building of category 2 once, a building of category 3 zero times, a building of category 4 three times, and a building of category 5 zero times.

The categorical vector loses the dimension of time within a single day, because we no longer track the order in which the participant visited the buildings, while the time series representation keeps this information intact. Losing a dimension of time within one day may be beneficial because it allows higher edge connections between similar behaviors no matter what time they occur at, for example a participant who visits the gym in the morning would remain highly connected to another participant who visits the gym in the evening. However, losing the dimension of time within a single day may also lower the graph's performance because it removes information that may be useful. We will test metrics that use each of these representations to see which one performs better.

2.4.2 Euclidean Similarity Euclidean Similarity is a metric for finding the similarity between two categorical vectors. We define the Euclidean Similarity between two categorical vectors X and Y as:

$$EuclideanSimilarity(X,Y) = \sum_{i=1}^{5} \frac{5}{2} X_i Y_i$$

This metric gives us the similarity between any two patient's categorical vectors, which thereby gives us the level of similarity for any two patients per day. This similarity metric is simple in nature, however that comes at a trade-off. The trade-off with euclidean similarity is that it does not account for time of day when comparing two trajectories. Because it uses categorical vectors it loses the dimension of time within a day's data points.

2.4.3 Levenshtein Edit Distance Consider defining distance as the difference between two categorical trajectories. The trajectories are sequences of digits (1-5) so this problems boils down to getting the distance (or similarity) between two sequences.

This problem of sequence comparison is one that the bioinformatics world has many metrics for. There are many useful metrics for comparing two sequences, one of which being string edit distance.

String edit distance is one of the most popular sequence comparison techniques. The algorithm results in the Levenshtein distance between two sequences a and b, which is the equal to the minimum number of insertions, substitutions, and deletions required to transform string to string to the other. More formally, Levenshtein edit distance between two string a and b is defined as:

$$\mathrm{lev}_{a,b}(i,j) = \begin{cases} \max(i,j) & \text{if } \min(i,j) = 0, \\ \min \begin{cases} \mathrm{lev}_{a,b}(i-1,j) + 1 \\ \mathrm{lev}_{a,b}(i,j-1) + 1 \\ \mathrm{lev}_{a,b}(i-1,j-1) + 1_{(a_i \neq b_j)} \end{cases} & \text{otherwise.} \end{cases}$$

Fig. 1. Definition of Levenshtein edit distance between two string a and b. We define lev(i,j) as the edit distance between a[1...i] and b[1...j]

Edit distance is a local sequence comparison metric. This is because edit distance considers alterations to a specific character without considering swaps from other parts of the string. The implication of this feature is that edit distance does not account for sequences which may be common over time however not in a specific location. For example, consider the sequences 22221 and 12222. The edit distance between these two strings is 2 because there are 2 substitutions required to transform one string to the other, one for the first character and one for the last character. This edit distance algorithm cannot recognize that a swap of the first and last character would make the string equivalent, thereby requiring no substitutions. Thus, the algorithm only picked up on local changes required, without considering the global structure of the string.

We have participants who may visit a category in the morning while other participants visit the same categories in the evening, like the gym for example. These sequences, based on this edit distance algorithm, would not be recognized as similar due to the difference in times at the gym. Thus, this may be a flaw in the metric, however we will objectively test it against other metrics to see how it performs.

2.4.4 Global Sequence Alignment The global sequence alignment technique addresses the weakness of local edit distance, because it considers the global nature of a trajectory. This metric takes two strings as input and outputs a similarity score between the strings based on how many operations need to be applied to transform one string to the other, similar to edit distance. However, the key advantage for global sequence is the ability to swap characters across the sequence. The global sequence alignment algorithm accounts for swaps in characters to take into account the global nature of the string. This should, in theory, handle the problem of two participants going to similar locations at different times of day. We will analyze the performance of this technique in the results section. An important thing to note is the character swaps that this algorithm allows are single swaps, not groups of characters, and that the operations can have variable weights to them. Thus, the operation weights are parameters that can be tuned.

$$\text{score} = \max \begin{cases} F(i-1, j-1) + s(x_i, y_i) \\ F(i, j-1) - \text{gap penalty} \\ F(i-1, j) - \text{gap penalty} \end{cases}$$

$s(x_i, y_i)$ may be +1 or −2
depending on match/mismatch

Fig. 2. Definition of Global Alignment between two string x and y, using the Needleman Wunsch algorithm. We define F(i,j) as the global alignment distance between x[1...i] and y[1...j]

As you see, the operation costs $s(x_i, y_i)$ are variable parameters that can be tuned. Also, the gap penalty is variable, which represents the penalty to the alignment for a case of ”swapping” characters across the string. In the results section, we will discuss the different parameter tuning sets we attempted to optimize the accuracy metric, and their respective results.

2.4.5 Pair-Wise Alignment Global alignment has one major drawback which pair-wise alignment attempts to tackle. The issue with global alignment is that is requires that every operation that transforms one sequence into another be done on a single character in a sequence. For example, in the sequence $a = [1, 3, 2, 1]$ compared to the sequence $b = [2, 1, 1, 3]$, the global alignment algorithm would incur a gap penalty for swapping $a[0]$ with $a[2]$, and another penalty for swapping $a[1]$ with $a[3]$. The pair-wise sequence alignment algorithm would recognize that a more efficient swap can be done pair-wise, where $a[0 : 1]$ can be swapped for $a[2 : 3]$ which limits the overall number of swaps, thereby only incurring a single gap penalty. Pair-wise sequence alignment has the ability to recognize opportunities of group swaps rather than only single character swaps, which causes for less gap penalties when there are more local similarities in clusters across a sequence.

The physical significance of this ”pair-wise swap” is that it accounts for similarities in trajectories that exist in local clusters. For example, consider a patient A who has records in their residence hall, then the gym, then in a classroom, then back to their residence hall and another patient B who has records in a classroom, then gym, then a residence hall for two consecutive records, and finally the gym. For the sake of example, let’s represent their trajectories as $A = [R, G, C, R]$ and $B = [C, R, R, G]$. In the global alignment algorithm, the distance between A and B would incur a penalty due to 2 swaps, namely $A[0]$ with $A[1]$ and $A[2]$ with $A[3]$. However in the pair-wise sequence detection algorithm, this incurs a single swap penalty with a swap of $a[0 : 1]$ and $a[2 : 3]$. The advantage of this single penalty is that these two trajectories are relatively similar, the only difference is student A had class in the middle of his/her trajectory while student A had class in the beginning of his/her trajectory. Thus, to account for different schedules amongst students, which is very common in the student population, the pair-wise algorithm better matches the necessary groups. These two students still visit then gym and have the the same number of residential and classroom time, thus the distance between the two should be minimal - which pair-wise alignment offers.

This approach can be tuned similar to global alignment, because the gap penalties and operation costs are variable. We will test different combinations of parameters to see which results in the best accuracy. Pair-wise alignment models our physical constraints the best because it can account for differences in students schedules without sacrificing the similarity between key behaviors (like going to the gym or spending excess amount time in a residence hall). We will show the results for pair-wise vs the other techniques described above in the results section.

2.5 K-Neighbors and Spectral Clustering to Determine Distance Metrics

Now we've seen many possible distance metrics to define the edge weights in the graph, the question of comparing each metric is important. How do we determine which metric is most fitting for our data set? The edge metric function determine the structure of the graph, thus it must be done with careful considerations because the structure of the graph can drastically change the performance of the GNN model.

We knew that the nodes in the graph represented patients, because our unit of classification is a patient. Therefore, the decision of edges, or how to connect said nodes, was one we had to consider most. There are a plethora of distance metrics to connect patients in our graph, ranging from physical features, spatial features, to social features. We had to objectively decide the best metric for connecting patients in our graph. This led to a pipeline that would objectively score each metric, to give us a means of picking the best metric. We decided to use a common practice in the GNN community, which is a K-Nearest Neighbors clustering algorithm to compare each metric.

This pipeline would start with developing a graph using a given distance metric. This would result in a graph with distances (or similarities) between every single node (i.e. participant). We then reduce the fully-connected graph based on the well-known K-Nearest Neighbors algorithm to capture only the K neighbors that are most similar based on the metric for every node. Once the K-Nearest Neighbors graph is created, it is sent to a clustering algorithm, specifically Spectral Clustering. Spectral Clustering is an unsupervised learning technique that uses the structure of the graph to classify each node into one of two clusters, depressed and non-depressed. Then, these clustering labels are compared to the true clinical labels through a final F1 score. This F1 score represents how well the given metric performed on clustering the data set. Figure 3 depicts this pipeline and its components.

By comparing the F1 scores produced by each metric, we can objectively determine which metric will perform best in the GNN. Now let's look at some of the metrics that we analyzed.

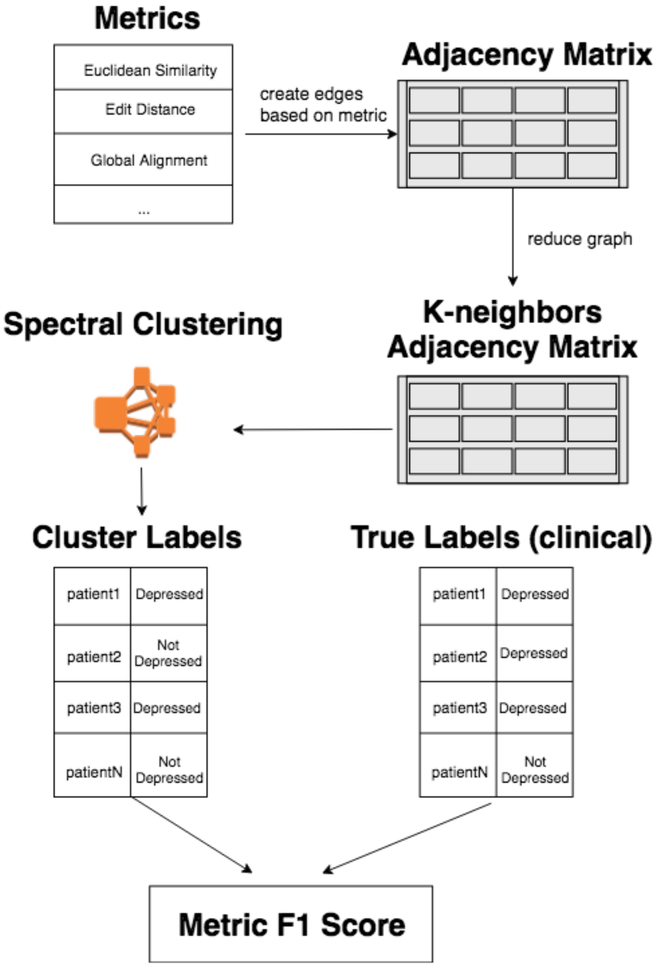

Fig. 3. A pipeline to determine the best distance/similarity metric, using K-Nearest-Neighbors and Spectral Clustering.

3 Results and Discussion

3.1 Fitbit Correlations

First, we tested the correlation of Fitbit sensing markers and depression. This correlation analysis was to ensure our graph could use Fitbit data as the node representations for patients. If the Fitbit data is correlated with depression, then we could utilize said data in each node for our graph representation.

To test the correlation of Fitbit sensing data with depression, we broke down the analysis to see the correlations by feature. We were interested in the correlations for each feature with two labels - the clinical results and the QUIDS survey results. We wanted to analyze both of these results, because this gives us a multi-view approach to feature selection, meaning we can gain more insight into which features are most correlated with depression. We calculated p and r values for each feature vs each label to indicate how well each feature correlates with depression. Table 1 shows the results.

This analysis yielded high correlations between the three Fitbit categories (activity, sleep, and heart rate) and the two types of labels. Activity is captured by some features such as the number of active minutes, the number of floors climbed, the amount of burned calories, the amount of time spent burning calories, and the

Feature	Label	p-value	r-value
lightly_active_minutes	clinical_result	0.050127359	-0.050127359
floors_climbed	clinical_result	10^{-10}	-0.093226406
minutes_to_fall_asleep	clinical_result	0.058858266	0.028588001
OOR_hr	clinical_result	0.433588509	0.0118513
sleep_efficiency	clinical_result	10^{-27}	-0.162609758
total_time_in_bed	clinical_result	10^{-26}	-0.157873758
restless_duration	clinical_result	10^{-11}	0.102119318
Burn_cal_out	total_quids	10^{-11}	-0.10449499
Burn_minutes	total_quids	0.010746282	-0.038592674
sedentary_minutes	total_quids	10^{-6}	-0.07112363
total_distance	total_quids	10^{-6}	-0.070788748
floors_climbed	total_quids	0.000400232	-0.053538494
very_active_minutes	total_quids	0.011413923	-0.038274002

Table 1. Correlation statistics (p and r values for α = 0.05) of Fitbit features and two different types of labels: clinical and QUIDS score.

total distance walked. These activity features, amongst others not shown above, are very correlated with both clinical results and total QUIDS score. Thus, activity is clearly related to depression - which aligns with past depression work. Sleep also highly correlates with the labels we tested, as you can see by the minutes to fall asleep, sleep efficiency, total time in bed, and restless duration statistics. Finally, heart rate also correlates with depression, but not as strong as the other categories. The strongest heart rate indicator was out-of-range (OOR) heart rate, which is the heart rate which is out of the ”normal” range for a Fitbit user (too high or too low). This feature captures high variance in heart rates, which is a symptom of depression. The other heart rate features were not very highly correlated with the two labels, most likely because heart rate tracking has lots of potential errors. For example, some people wear their Fitbit at different parts of the day, which alters their daily heart rate statistics. Heart rate sensing is still a developing technology in Fitbit devices, thus has more error margins than the other categories (sleep and activity). That being said, some heart rate measures, like OOR heart rate, have a high correlation with depression labels.

The conclusion from this first step is that Fitbit activity, sleep, and heart rate data is correlated with depression. This is true for two depression labels, clinical results and QUIDS scores. This analysis also yielded the specific features that are most relevant to our study, because some features (like sleep efficiency and total burned calories) are much more correlated with depression than others. This gives us insight into which features our neural network will need to weight the most.

Now we’ve observed that Fitbit data is significantly correlated with depression labels, so we can use Fitbit data to build the vertices in our depression graph. The graph will represent each patient as a vertex, because the unit of the vertices is what a GNN classifies. Each patient is represented as his/her set of Fitbit features, which is representative of their activity, sleep, and heart rate data. This is a valid approach to building the graph because we’ve proved that Fitbit data correlates with depression levels, which is what we aim to predict. Now that we have the vertices defined in the graph, we must determine how to define edges.

3.2 Edge Metric Analysis

Edges represent the connections between individuals in our graph. An edge between patient A and patient B must signify that patients A and B exhibit similar behavioral trends of some sort, and are likely to have the same classification result. In the methods section, we defined a few metrics for defining the distance between two patients. In this section, we will dive into the results of each of those metrics to see which performs the best, and should eventually be used when constructing the graphs.

3.2.1 Heat-map Analysis Most of the presented metrics used GPS data to define the distance between patients, but before showing the results of those metrics we will show why we chose to use GPS metrics. The decision to use GPS data as metrics in edge distances was not trivial, but rather was a result of visual and statistical analysis of our data set. The first way we confirmed that GPS data was a strong candidate for edge distances was through a visual analysis tool called heat maps. Heat maps are spatial maps that track how popular each area in the map is, using a dimension of color. The ”heat” of any given part of a map is defined by the intensity of the color that lays on it. The ”hotter” a section of a map, which means the more intense the colors in a section are, the more popular that section in the map is. By making heat maps for both depressed and non-depressed individuals, and visually comparing the two sets of maps, we could understand the different spatial patterns of both groups. The goal here is to observe differing behavior in the two groups so we can prove that GPS data is a good indicator for edge distance. Shown in figures 4 and 5 are one set of such heat maps, normalized for the number of people in both for a specific day (March 3, 2017).

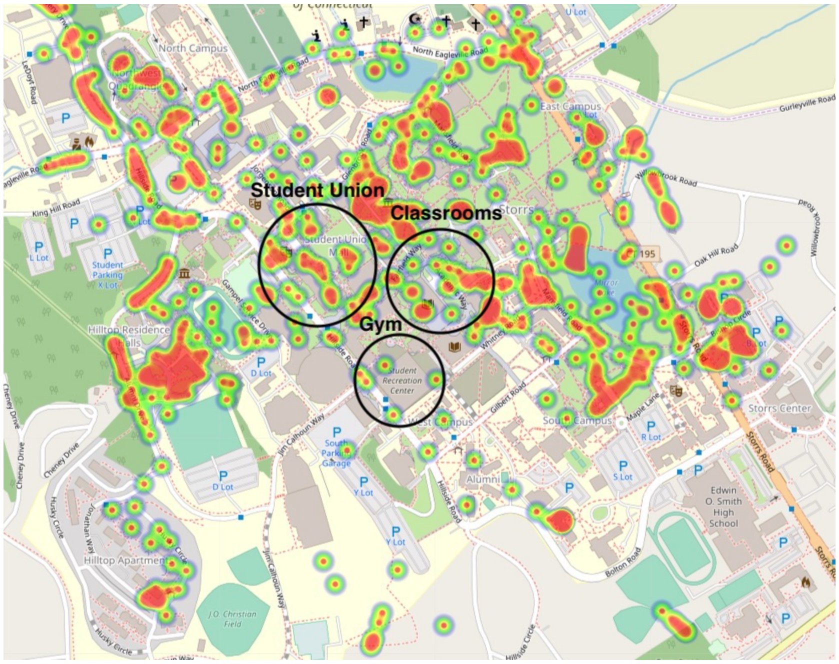

Fig. 4. Heat map of depressed patients on UConn’s campus from 3/3/17, with circles to bring attention to three specific locations.

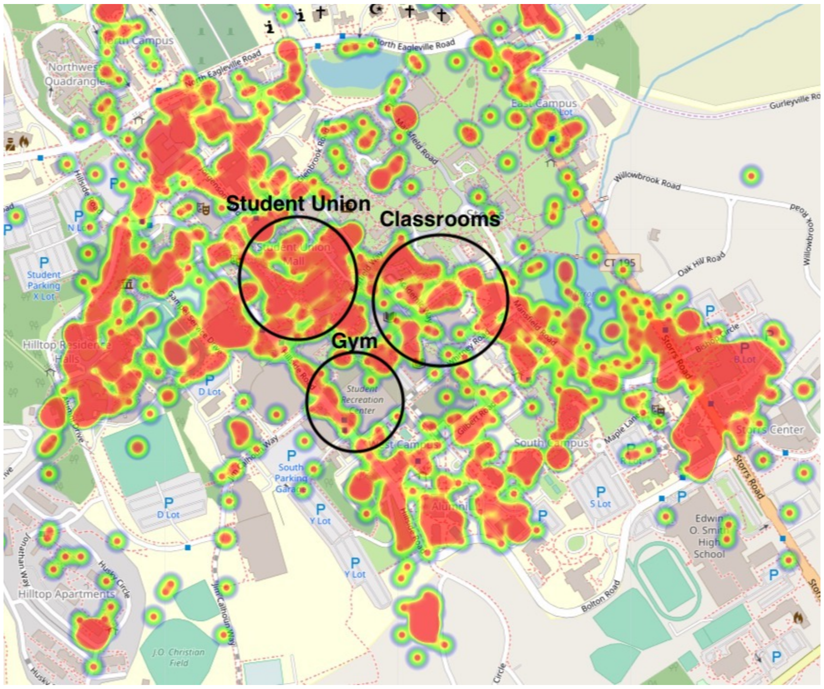

Fig. 5. Heat map of non-depressed patients on UConn's campus from 3/3/17, with circles to bring attention to three specific locations.

These two figures show many insightful details about the geo-spatial movements of the depressed and non-depressed subsets. At first you'll recognize the density of the non-depressed map is more, which is because the non-depressed group has more movement around UConn's campus. This aligns with the reason why activity is correlated with depression. Beyond that, there are three specific locations in both maps that are circled because those locations convey more details.

The first location is the Student Union. The Student Union (SU) is the social center of UConn's campus, because it is located in the center of campus and is where most students eat lunch and participate in extracurricular activities. There are shows and movies regularly, an arcade, pool table, a bowling ally, and other entertainment facilities there. It is UConn's largest entertainment area. You'll notice how the depressed group has very few points in the SU compared to the non-depressed group. This indicates that depressed patients are not visiting these entertainment spots as often as non-depressed students. This aligns with the mental nature of depression, because those who do not visit entertainment locations are more prone to build up mental fatigue and stress.

The second location is the Student Recreation Facility, or the gym. The gym is an indicator of physical activity. Again you'll see many more data points in the non-depressed case in the gym compared to depressed, which aligns with the Fitbit correlations. In the Fitbit correlations, we observed that high levels of activity correlated with low levels of depression, thus more time spent at the gym should be seen in the non-depressed group because the gym is a place for activity.

Lastly, the third location is the main classroom area on UConn's campus (a combination of Laurel Hall, Oak Hall, and the others). This area is also denser in the non-depressed heat map, which indicates that more non-depressed students are going to class than depressed. This aligns with the depression symptom of limited social exposure. By not going to class, the depressed students are missing an opportunity to socialize with peers and talk to professors.

Based on a visual analysis of the three specific areas mentioned above, it is clear that GPS data has a correlation with depression. This is, however, not a rigorous enough correlation because it is visual - not numerical. Thus, we will now show the statistics that prove these differences in spatial patterns between the depressed and non-depressed groups.

Location		Depressed		Non-Depressed	
		# of points	# of unique users	# of points	# of unique users
Acadmic Campus	avg	287.5	6.083333333	564.4166667	14.91666667
	median	280.5	7	552	17
Union	avg	23.33333333	2.583333333	13.66666667	3.75
	median	10.5	3	11	4
Gym	avg	14.66666667	2.833333333	14.25	3.833333333
	median	16	2.5	9.5	3.5

Table 2. Location statistics for number of GPS points and number of unique users in the academic campus, Student Union, and gym over a two week window.

With the numbers from Table 2, it becomes numerically obvious that there are, on average, more non-depressed students in the Union, gym, and on the academic campus. The area where this is most obvious is definitely on the academic campus. The academic campus has on average 145% more non-depressed students than depressed students, after normalization. This implies that it is much more common for non-depressed students to attend classes than depressed. This difference is extremely significant, because it is clearly a large statistical difference. This implies that GPS data can be used to differentiate depressed and non-depressed individuals, because there are large differences in the coordinate patterns of each group.

3.2.2 Categorical Analysis Now that we've observed the fact that GPS data is a valuable asset in differentiating depressed and non-depressed people, we can dive deeper into how to use GPS data in building edge connections. One initial thought was to use raw GPS points as a means of connecting people in the graph (i.e. building edges), however this is problematic because people can visit different locations that serve the same purpose. For example, a student who visits the Chemistry Building for most of their day is likely attending classes, and a student who visits the Information Technology & Engineering Building (ITEB) is also likely attending classes - however raw GPS points would show these two students as very far apart. Also, students live in many difference areas, thus raw GPS points would show high distances between people who spend their time in residence areas, but we want those types of students highly connected because it implies they are not socializing and are not active. Thus, raw GPS data was not a good edge metric.

Being that this data was collected on UConn's academic campus, we have the advantage of knowing which buildings people could potentially visit throughout the study. Thus, we categorized each building, as explained in the Methods section, to categorize each GPS data point. The categorization addresses the issues that arise when using raw GPS data because we could build connections based on categories, so individuals attending classes would be highly connected, individuals spending excess time in residence halls would be highly connected, etc. Before we move forward with using categorical GPS data, we needed to observe that the data shows a clear divide between depressed and non-depressed individuals. This difference must be statistically significant because it will be the basis of the edges in our graph. Thus, we tested this approach to see what are the categorical differences between the depressed and non-depressed groups. The data in Table 3 is normalized using random shuffling, to ensure there are the same number of participants being compared.

Average Points Per Day			
Category	**Purpose**	**Depressed**	**Non-Depressed (normalized)**
1	Entertainment	64	142.92307
2	Sports	12.53846	12.3846153
3	Department	100.846153	133.07692
4	Residence	393.15384	273.30769
5	Library	11.92307	8.153846

Table 3. Average number of categorical GPS points from 2/26/17 - 3/10/17 for depressed / non-depressed groups.

Table 3 shows large differences in the number of categorical points for the categories of entertainment (79% difference), department (28% difference), and residential (36% difference). These differences align with what we've observed thus far, which is that depressed people spend more time in residence areas, less time in entertainment areas, and less time in classrooms. The other two categories, sports and library, are not significantly different most likely because there are not many data points in those categories. There is still a difference, as you'll see there are more depressed data points in the library than non-depressed, but the sample size is small.

Thus, these heat maps and categorical statistics prove that categorical GPS data is a valid signal to separate depressed vs non-depressed. This allows us to move forward in using categorical GPS data for edge metrics. Now the question is how should we use categorical GPS data to define the distance between individuals in our graph.

3.2.3 Distance Metric Results In the Methods section, we defined a few distance metrics that leverage GPS categorical data to build edge connection in the graph. Those metrics were: Euclidean Similarity, Levenshtein Edit Distance, Global Sequence Alignment, and Pair-Wise Alignment. Each of these metrics were tested using the metric pipeline depicted in Fig. 3. The results of said pipeline is shown in table 4.

Table 4 makes it clear that pair-wise global alignment, specifically when k=3, has the best performance in spectral clustering. This makes sense because pair-wise global alignment has all the features of global alignment plus the advantage of recognizing pairs of swaps to make. The Methods section explains this advantage in depth.

Thus, we now have a distance metric that is the statistically best of the options. Spectral clustering used only the GPS features to classify with 69% accuracy, so we aimed for our GNN to have a higher accuracy than plain spectral clustering. We can now define our graph as follows: the vertices in our graph will be represented by Fitbit feature data, which was derived from the results of section 3.1. The edges in our graph will be represented as distances between categorical trajectories, which was derived in results from sections 3.2, specifically using pair-wise global alignment to define the distance between trajectories. These decisions were all rooted in the results shown above. Now we explain the performance of GNN on this graph.

Distance Metrics Spectral Clustering Accuracy							
euclidean similarity		**edit distance**		**global alignment**		**pair-wise global alignment**	
k	F1 Score	k	F1 Score	k	F1 Score	k	F1 Score
2	0.5710822291	2	0.6509465511	2	0.5468620373	2	0.6932717119
3	**0.5806007405**	3	**0.6626695706**	3	0.5266306452	3	**0.6932717119**
4	0.5710822291	4	0.6626695706	4	0.6011411785	4	0.5585970684
5	0.5777587204	5	0.6396515984	5	**0.6171110376**	5	0.5828821699
6	0.5710822291	6	0.5666290815	6	0.5341054193	6	0.563181671

Table 4. Spectral clustering accuracy for each distance metric, given different k parameter in K-Nbrs Graph

3.3 Graph Neural Network Performance

The graph was built using patients' Fitbit features as the vertices and categorical GPS differences as edge distances (namely, pair-wise global alignment). We built a k-nearest neighbors graph to remove any need to set and tune a threshold distance for defining edges. We built a graph for every day over the course of our study, where the Fitbit features changed for each patient based on their daily features and the edge metrics were computed for everyday based on daily trajectories.

Once the graphs were built, we had to run a GNN model on those graphs. There is a plethora of available literature and tools for GNNs, so we found a popular PyTorch library to handle the GNN for us. We built the graph with 2 layers, both GraphConvolution layers. The first layer took as input 74 floats, one for each feature, and reduced those features to 16 to be passed to the next layer. This was chosen through tuning of hyper parameters. The second layer was also a GraphConvolution which took as input 16 inputs and had 2 output signals, one for the probability of class 0 and another for the probability of class 1. The activation function was a logarithmic softmax, which provides the probabilities for each of the two classes. This is a popular activation function in lots of past GNN literature.

Overall, the model was simple - only 2 convolutional layers. This is because our dataset is relatively small, and a larger neural network was prone to overfitting very quickly. This simplistic model suited our use case well.

The data was divided into 65% for training, 20% for testing, and 15% for validation. This was also tuned, to find a combination which led the network decreasing loss without overfitting. The models were run for a varying number of epochs, but based on tuning 100 epochs was always enough to get loss to a minimum.

Table 5 shows a subset of the results, those over days with the most number of data points available. As you can see, the GNN performed very impressively compared to the simple spectral clustering benchmark (69% vs 85%). A basic classifier would give 50% accuracy, and a basic model (like spectral clustering) could give no more than 70% accuracy. We beat all those models by at least 15% accuracy, which shows the advantage of using a GNN. Thus, we observed that GNN performed extremely well given our graph - confirming the approach's dominance in node classification problems.

Date	acc_train	acc_val	test_acc
3/1/17	0.913	0.8	0.8571
3/2/17	0.8889	0.78	0.875
3/3/17	0.8889	0.75	0.875
3/4/17	0.9048	0.8	0.7143
3/5/17	0.8	0.741	0.875
3/6/17	0.8966	0.7143	0.8889
3/7/17	0.9062	0.75	0.9
3/8/17	0.8519	0.7143	0.875
3/9/17	0.871	0.7143	0.8889
3/10/17	0.7692	0.78	0.75
avg:	**0.86905**	**0.75439**	**0.84992**

Table 5. GNN training, validation, and test accuracies throughout 10 days.

4 Conclusion

4.1 Learnings / Insights

This analysis project has uncovered a few key pieces of information. First, we statistically confirmed hypothesis 1 by showing that certain Fitbit features are highly related to depression levels. Second, we observed that GPS categorical data is very valuable in separating depressed and non-depressed. This was confirmed through visual analysis (heat maps) and again through statistical analysis. Finally, we confirmed hypothesis 3 that GNN performs very well on node classification problems such as depression. Depression exhibits many characteristics, physical and social, which allows a graph representation to capture more dimensions of data than a normal data set representation. Therefore the GNN approach, when used on a well-devised graph, proved to be a superior method of classification (over simple heuristics and spectral clustering). This model gives insight into the disease of depression, and how it manifested in our data set, but also insight into the power of GNN.

4.2 Future Work

This study is helpful in showing a novel machine learning technique, graph neural networks, are very useful in the domain of depression prediction. One insightful addition would be to investigate which features the neural network highlighted in its classification, to understand which features are most relevant when diagnosing depression. Doctors would benefit from statistical models and features to look for in depression candidates. In the future, this prediction model can go beyond just predicting by giving insight into what signals make good predictions to assist doctors in their own diagnoses.

One major constraint of our study was the ages of the participant pool. All participants were college students, so we did not have a very diverse age range. Future work should be done to investigate similar trends in other age ranges, to find features that are most relevant to other age groups.

Another area of improvement is better utilizing the QUIDS survey results for classifications. In the next iteration of this study, we will build a temporal graph approach which uses multiple graphs to predict a single label per week, because we have Fitbit data per day and QUIDS results per week. Thus we can build a $multi - view$ approach where each day is its own view on a given label. This leverages the QUIDS data, and can be combined with the model built in this study for an additional view.

References

1. Depression: clinical, experiental, and theoretical aspects, 1969
2. Depression statistics, 2019.
3. A. T. Beck. *Depression: clinical, experimental, and theoretical aspects*. Staples Press, 1969.
4. P. Berto, P. Ruffo, and R. Virgilio. Depression: costofillness studies in the international literature, a review. *The Journal of Mental Health Policy and Economics*.
5. F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini. The graph neural network model. *IEEE Transactions on Neural Networks*, 20(1):6180, 2009.
6. H. D. Schmidt, R. C. Shelton, and R. S. Duman. Functional biomarkers of depression: Diagnosis, treatment, and pathophysiology. *Neuropsychopharmacology*, 36(12):23752394, 2011.
7. M. Simonovsky and N. Komodakis. Dynamic edge-conditioned filters in convolutional neural networks on graphs. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.