Machine Learning Final Project Spring 2023

Ioana-Andreea Cristescu
Department of Computer Science
University of Richmond
Richmond, Virginia
ioanaandreea.cristescu@richmond.edu

Ying Zhu
Department of Computer Science
University of Richmond
Richmond, Virginia
ying.zhu@richmond.edu

Shewta Ware
Department of Computer Science
University of Richmond
Richmond, Virginia
sware@richmond.edu

I. Minimum Requirements of Project

We are using approach 2 with the following requirements: Data cleaning, Data visualization, Feature extraction, Correlations, and Supervised learning.

II. Project Title

Predicting ADHD Using Smartphone Sensing Data

III. Project Aim

In this project, we will explore a classification problem for ADHD prediction. We use sensing data collected through an Android/iOS smartphone application to predict ADHD in a small pilot sample. This approach is based on continuous behavioral data collected automatically from participants' phones. We then use machine learning to predict if a patient has ADHD or not.

IV. Overall Plan

A. Introduction

ADHD is an impairing disorder that affects multiple domains of functioning and manifests itself in the daily lives of different age groups with the disorder. To help people prevent the negative impact of ADHD, we conduct research with college students, using machine learning models and sensor data to predict if the individual has ADHD or not.

B. Related Work

There are numerous studies that employ smartphone sensing data in assessing and detecting mental health diseases such as ADHD and depression [1], [2], [3], [4]. In the following, we take inspiration from related work and create something novel.

A previous study used smartphone sensing data collected through an Android app in order to predict ADHD symptoms [2]. Ware et. al. (2022a) used SMS data collected in time intervals of 7 days along with their ADHD symptoms scores. In our work, we will be using data collected thorough both an Android and iOS app. Depending on how the data was collected, we will either analyze the training samples separately or in conjunction. Moreover, instead of using SMS data, we will utilize activity data recorded for each individual day or for a time interval that will be determined after performing data analysis. The ultimate goal of our project is to predict whether a patient has ADHD or not.
Ware et. al. (2022b) uses a sliding window approach in the attempt of capturing behavior patterns and changes. For example, for a day t, they would consider the data collected during the past n days, i.e., $[t-n+1, n]$. The value of n used in Ware et. al. (2022b) was $n = 7$ or $n = 14$. We will employ a similar approach for only $n = 7$ as each participant filled out a weekly survey. Moreover, we will attempt to improve the sliding window approach such that we are able to accurately capture the activity behaviour patterns of an individual.

By working with smartphone sensing data that is susceptible to various unpredictable variables, we observe significant amount of missing data. Yue et. al. (2021) presents an approach of dealing with GPS missing data collected through Android and iOS apps. Similar to their findings, our missing activity data can happen during the day or night due to scheduling of the operating system, failure of data capture by sensors, or mis-configuration by a participant. Using this information along with our own personal knowledge about the schedule of a college student during the week days and weekends, we will be able to develop an efficient way of substituting or discarding missing data.

C. Dataset Used

With consent, we have gathered sensor data from the app designed for both Android and iOS users, to cover the majority of the mobile phone choices of college students. The sensor data include phone call data, message data, WiFi location data, activity, and app usage data. In this stage of the research, we will focus on the activity data. The data marks the sensor starting time, sensor type, activity, and confidence in the data. The ”activity” element is described as ”still”, ”tilting”, ”running”, etc., corresponding to related human activities.

Besides the sensor data, we also have the self-report questionnaire answered by each participant. Within the self-reported surveys, we list 18 questions about inattention, hyperactivity, and impulsivity. Accompanied by the official diagnosis from physicians, we can further use the self-reported data to find correlations between sensor data and self-reported data later in the research.

D. Approach

The approach we took in conducting this research is comprised of four stages as illustrated in Figure 1. After recruiting the participants with both Android and iOS devices, they were instructed to fill out an initial questionnaire about their ADHD diagnoses and symptoms, followed by weekly surveys on inattention, hyperactivity, and impulsivity. At the same time, the participants were instructed on downloading the Android or iOS app for recording their activity. In the third step of our research, we pre-process the data, calculate an array of features, and use various machine learning models for ADHD prediction. In the fourth, and final step, we predict on unknown data by utilizing the models trained in the previous step.

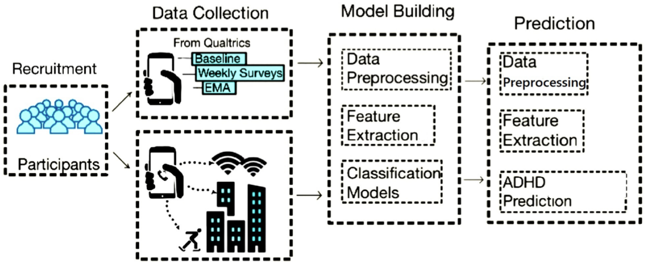

Fig. 1. Illustration of high-level approach.

E. Data Analysis Steps

1) Data Cleaning: We will first clean the data collected by removing the duplicates and filtering it based on the start and end date. In this research, we will only focus on a certain time span of the participants' activities between 10/14/2022 and 12/30/2022.

2) Data Visualization: We will visualize the data collected by creating a histogram plot. The plot will represent the frequency of the number of days between 2 self-reported surveys filled out by all participants. The x-axis will represent the number of days, such as 1, 2, 3, 4, ..., and 7 days. The y-axis will constitute the number of times the specific time interval occurred. With this plot, we can best select a threshold and discard the data that is below the threshold. With the remaining data, we will create another histogram plot with the sensor data between 2 self-reported surveys. The aim is to find the optimal time interval value to group the sensor data. We will use the number of days as the x-axis, from 1 to 7 days, and the number of times as the y-axis. Then, we will select the most appropriate time interval from observing the histogram.

F. Methodology

1) Data Scaling and Feature Extraction: After determining the most appropriate time interval from the previous steps, we will focus on feature extraction based on the certain time interval n. We will extract features such as activity type, activity duration during one day, activity duration during n days, etc. The activity type can be further detailed into "in vehicle", "on bicycle", "running", "still", "tilting", "walking", and "unknown". We will further analyze the data labeled with "unknown" for potential data grouping or discarding.

Then, we will employ min-max scaling on the quantitative data.

2) Models: We will try out several classification models to determine the ones that will return the best results. Some classification models will include SVM, Random Forest, and XGBoost. Detailed parameters and model tuning will be determined in the experiment process.

G. Expected Results

After training the SVM, Random Forest, and XGBoost to classify the presence of ADHD in a patient, we will compare their performance by calculating the F1 score. We anticipate that the stationary time during one day, as well as during n-day time interval will be negatively correlated with the ADHD diagnoses, while the time spent performing activities such as running or riding a bike will be positively correlated to whether an individual has ADHD or not.

V. Tentative Timeline and Responsibilities

Week of 3/19: Data cleaning; data visualization on both the survey data and sensing data; mapping ADHD label and total ADHD score

Week of 3/26: Data cleaning (missing data); Feature extraction and scaling; ready to run the model for 2 different time intervals

Week of 4/2: Gather data from the models; calculate expected results (F1, correlation, etc.)

Week of 4/9: Further analysis of the data gathered; Plot making for report

Week of 4/16: Finishing the report and presentation

VI. Optional-Above and Beyond

We will explore Pearson Correlation between the ADHD self-reported scores and features we used in the model. The correlation will measure the strength of the linear relationship between the two variables above.

We will explore other types of data collected by the Android/iOS apps such as location, app usage, and phone call/message data.

References

[1] R. Wang, F. Chen, Z. Chen, T. Li, G. Harari, S. Tignor, X. Zhou, D. Ben-Zeev, and A. T. Campbell. Studentlife: assessing mental health, academic performance and behavioral trends of college students using smartphones. In *Proceedings of the 2014 ACM international joint conference on pervasive and ubiquitous computing*, pages 3–14, 2014.

[2] S. WARE, L. E. KNOUSE, I. DRAZ, and A. ENIKEEVA. Predicting adhd symptoms using smartphone sensing data. 2022a.

[3] S. Ware, C. Yue, R. Morillo, C. Shang, J. Bi, J. Kamath, A. Russell, D. Song, A. Bamis, and B. Wang. Automatic depression screening using social interaction data on smartphones. *Smart Health*, 26:100356, 2022b.

[4] C. Yue, S. Ware, R. Morillo, J. Lu, C. Shang, J. Bi, J. Kamath, A. Russell, A. Bamis, and B. Wang. Fusing location data for depression prediction. *IEEE transactions on big data*, 7(2):355–370, 2018.