Depression Group Prediction with Passive Sensing: Prediction Accuracy and Window Sizes

Sabinakhon Akbarova, Youngtae Noh*

sabina@nsl.inha.ac.kr. ytnoh@nsl.inha.ac.kr*
Department of Electrical and Computer Engineering, Inha University, Incheon, Korea

Abstract

Depression diagnosis is an important issue, usually requiring mental health professionals' help, which is not affordable to everyone in terms of time and money. Passive sensing is an emerging technology allowing effortless depression detection with everyday mobile devices. One of the main challenges in eHealth applications is the improvement of prediction accuracy. In this work, we investigate the impact of feature extraction window sizes on depression group classification accuracy. Our findings illustrate the importance of window size hyperparameter and the positive tendency of accuracy with the increase of window size. The highest accuracy that we could achieve for multi-class classification is 77% for the window size of 14 days.

Ⅰ. Introduction

Depression is a common and dangerous illness accounting for almost 300 million people affected worldwide [1]. Although depression symptoms manifest in day-to-day life, they are not straightforward to diagnose and assess. Seeking help from mental health professionals is the most effective method for thorough diagnosis and individually tailored treatment. However, this approach is time-consuming, high-priced, and requires professional involvement.

Previous studies demonstrated the potential of smartphone passive sensing in depression detection since sensor data provides meaningful insights about behavioral patterns that, in turn, are highly correlated with depression [2-6]. It is common to extract features from raw sensor data within several hours epochs [2,4,6] or aggregate them into daily measures [3,4,5]. However, previous works lack investigating the effect of window sizes on depression group classification. Therefore, we conducted a study that makes the following contributions:

- We developed a data collection agent and passively collected sensor data from 60 people for two months to explore the feasibility of smartphone sensor data to detect depression severity.
- We predicted the depression severity with various window sizes for features extraction and showed the importance of the choice of the window size parameter.

II. Study design

Figure 1 illustrates the overview of the study procedure. The first step is participants recruitment and data collection. After data collection is finished, features are extracted from raw data and preprocessed for further analysis. As the last step of the experiment, we feed data to the machine learning classifiers and test the resulting models on unseen data.

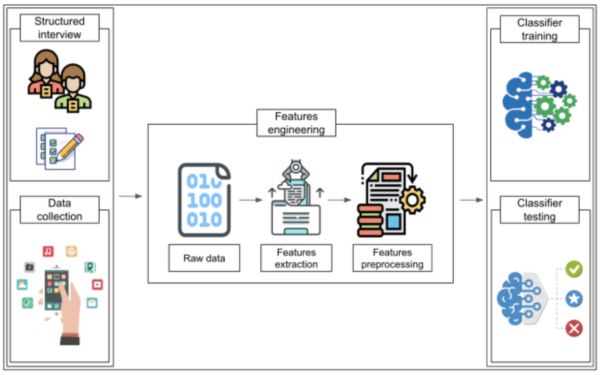
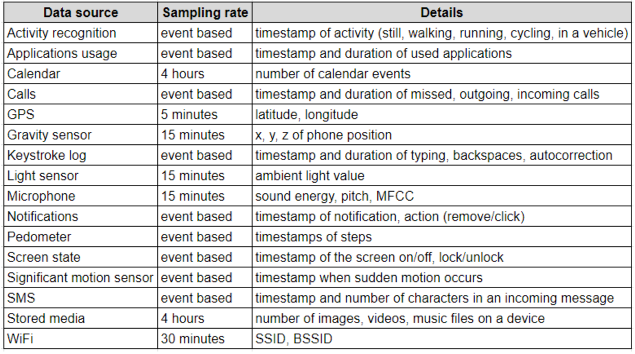


Figure 1. Study design overview

III. Data collection

The recruitment happened through online social communities. We recruited the participants on a first-come-first-serve basis. A total of four depression screening tools were conducted for depression severity assessment - Patient Health Questionnaire (PHQ-9) [7], Structured Clinical Interview for DSM-5 (SCID-5), Beck Depression Inventory-II (BDI-II) [8], and Korean Inventory of Depressive Symptomatology (K-IDS) [9]. Every applicant who wanted to participate in our study was required to, firstly, take a PHQ-9 online via Qualtrics [10]. We classified applicants into three groups (non-depressed, depressed group, severely depressed group) following the results of the PHQ-9 screening tool with cut-off scores of 10 and 20 (0-9 - non-depressed group, 10-19 - depressed group, 20-27 – severely depressed group). For the further pre-assessment, interviewers administered SCID-5 for depression severity confirmation because it is a well-established diagnostic assessment tool. The selected participants were required to fill in BDI-II and K-IDS tests online four times during the study with two weeks intervals.
As a result of the structured interview, we recruited 60 people (24 males, 36 females of average age = 25 years with a standard deviation of 6.1). Following the interview procedure described above, we identified three groups, namely non-depressed, depressed, and severely depressed, each comprising 20 participants. To verify the heterogeneity of the groups, we performed separate statistical tests for age and gender. We applied a chi-square test of independence to examine the relationship between gender and depression groups. The results (X2 (2, N=60) = 0.41, p>0.05) indicated no significant relationship between gender and depression group. The age characteristic is a continuous variable; and therefore, we used the Kruskal Wallis test, which resulted in p>0.05, indicating that the null hypothesis cannot be rejected and concluding that the group medians of age are equal for depression groups.
After being recruited, participants received the data collection agent "YouNoOne" and were instructed not to change the phone or uninstall the application. Data collection started only after a participant signed an informed consent form. The data collection duration was 60 days. We passively collected sensor data using the "YouNoOne" application. Table 1 summarizes sensor data collected with sampling rates and brief explanations. The collected data was securely transmitted to our cloud server via a TCP connection using SSL protocol [11]. The cloud server assigned random unique IDs for all participants to protect their identities.
The study procedure was approved by the Institutional Review Board of Yonsei University (No. 7001988-202103-HR-966-08)

Table 1. Collected data with sampling rates



Data source	Sampling rate	Details
Activity recognition	event based	timestamp of activity (still, walking, running, cycling, in a vehicle)
Applications usage	event based	timestamp and duration of used applications
Calendar	4 hours	number of calendar events
Calls	event based	timestamp and duration of missed, outgoing, incoming calls
GPS	5 minutes	latitude, longitude
Gravity sensor	15 minutes	x, y, z of phone position
Keystroke log	event based	timestamp and duration of typing, backspaces, autocorrection
Light sensor	15 minutes	ambient light value
Microphone	15 minutes	sound energy, pitch, MFCC
Notifications	event based	timestamp of notification, action (remove/click)
Pedometer	event based	timestamps of steps
Screen state	event based	timestamp of the screen on/off, lock/unlock
Significant motion sensor	event based	timestamp when sudden motion occurs
SMS	event based	timestamp and number of characters in an incoming message
Stored media	4 hours	number of images, videos, music files on a device
WiFi	30 minutes	SSID, BSSID

IV. Results

We performed data analysis after completing data collection. The data analysis pipeline starts from features engineering. We extracted more than 1500 features for different window sizes (around 300 features for each window size), namely a day, two days, three days, a week, two weeks. Our goal was to perform a comparative analysis of depression group classification accuracy for various window sizes. Each sensor has its own set of features, which we carefully selected based on related works. We calculated statistical characteristics such as mean, median, standard deviation, skewness, maximum, minimum, and contextual information (duration of homestay, phone usage duration at study places, etc.). The detailed list of features is out of the scope of this paper.

Features preprocessing pipeline consists of 3 steps: rare labels replacement, out-of-range values and outliers removal, missing data analysis. We treated labels that comprise less than 5% of the dataset as rare labels, which we combined into the "other" category. For example, we merged categories on a bicycle and in a vehicle into one. We treated duration features (call duration, moving duration, phone usage duration, etc.) that exceeded window size as out-of-range values and replaced them with missing data indicators. For outliers detection, we used 1% as a lower boundary and 99% as the upper limit, meaning that values falling out of borderline were considered outliers and regarded as missing data. Finally, we removed cases (rows of data) with more than 20% of missing values. We replaced the remaining missing values with the depression group median.

For the depression group classification, we decided to utilize tree-based ensemble algorithms, such as Random Forest (RF), AdaBoost, and XGBoost, because they do not require normal data distribution and are robust against overfitting. As hyperparameters, we used 100 decision trees in the RF classifier, 50 estimators for AdaBoost, and the maximum depth of 6 with a 0.3 learning rate for XGBoost. We split the dataset for train and test sets by 70% and 30%. We used GroupShuffleSplit from the scikit-learn library, where we treated each participant as a separate group. By doing so, we avoided having data from the same participant for training and testing sets. For model evaluation, we applied 10-fold cross-validation and calculated the accuracy as an average from the results. We repeated the experiment for all window sizes and summarized the results in figure 2. The highest achieved prediction accuracies are 68%, 73%, 77% for Random Forest, AdaBoost, and XGBoost, respectively. Figure 2 illustrates the generally positive prediction accuracy trend with the window size increase. Therefore, we conclude that by increasing the window size for features extraction, we can predict the depression group more accurately.

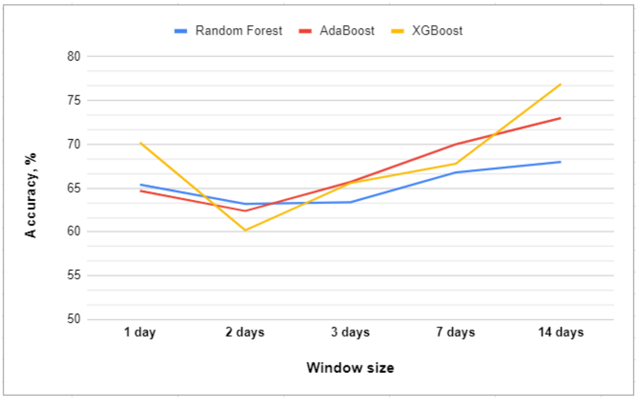


Figure 2. Depression group prediction accuracy for different window sizes

V. Conclusion

In this work, we investigated the influence of features extraction window size on depression group prediction. We collected sensor data from 60 participants from 3 depression groups (non-depressed, depressed, severely depressed) and extracted features from the data using different window sizes. Our results show that prediction accuracy generally increases with the increasing window size. In future work, we plan to experiment with more participants and more fine-grained window sizes.

ACKNOWLEDGMENT

This research was supported in part by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2019R1F1A1059898),

References

[1] World Health Organization. Depression 2020: https://www.who.int/news-room/fact sheets/detail/depression

[2] A. A. Farhan et al., "Behavior vs. introspection: refining prediction of clinical depression via smartphone sensing data," 2016 IEEE Wireless Health (WH), 2016, pp. 1-8, 10.1109/WH.2016.7764553.

[3] Rui Wang, Weichen Wang, Alex daSilva, Jeremy F. Huckins, William M. Kelley, Todd F. Heatherton, and Andrew T. Campbell. 2018. Tracking Depression Dynamics in College Students Using Mobile Phone and Wearable Sensing. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. 2, 1, Article 43 (March 2018), 26 pages. https://doi.org/10.1145/3191775

[4] A. Ghandeharioun et al., "Objective assessment of depressive symptoms with machine learning and wearable sensors data," 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII), 2017, pp. 325-332, doi: 10.1109/ACII.2017.8273620.

[5] Saeb S, Zhang M, Karr CJ, Schueller SM, Corden ME, Kording KP, Mohr DC. Mobile Phone Sensor Correlates of Depressive Symptom Severity in Daily-Life Behavior: An Exploratory Study. J Med Internet Res. 2015 Jul 15;17(7): e175. doi: 10.2196/jmir.4273. PMID: 26180009; PMCID: PMC4526997.

[6] Narziev, N.; Goh, H.; Toshnazarov, K.; Lee, S.A.; Chung, K.-M.; Noh, Y. STDD: Short-Term Depression Detection with Passive Sensing. Sensors 2020, 20, 1396. https://doi.org/10.3390/s20051396

[7] Kroenke, K.; Spitzer, R.L.; Williams, J.B. The PHQ-9: Validity of a brief depression severity measure. J. Gen. Intern. Med. 2001, 16, 606–613.

[8] Upton, J. Beck Depression Inventory (BDI). In Encyclopedia of Behavioral Medicine; Gellman, M.D., Turner, J.R., Eds.; Springer: New York, NY, USA, 2013; pp. 178–179.

[9] Journal of Korean Society for Depressive and Bipolar Disorders (우울・조울병), Vol.10(3): 131-151, 2012-10

[10] Qualtrics. Available online: https://www.qualtrics.com/

[11] Muhammad Salman, Touseef Javed Chaudhery, and Youngtae Noh. "Study on performance of AQM schemes over TCP variants in different network environments." IET Communications (2020).