Relating Physical Activity to Problematic Internet Use

Killian Conyngham*
Master of Data Science for Public Policy
Hertie School, Berlin, Germany
k.conyngham@students.hertie-school.org

Jackson M Luckey*
Master of Data Science for Public Policy
Hertie School, Berlin, Germany
jacksonmluckey@gmail.com

Fabian Pawelczyk*
Master of Data Science for Public Policy
Hertie School, Berlin, Germany
f.pawelczyk@yahoo.de

Abstract

The focus of our project is to predict the risk of problematic internet usage (PIU) by minors, based on data of their physical activities. In particular, our goal is to build a model for entrance in the Kaggle competition run by the Child Mind Institute on this topic. Here, one key challenge lies in the high level of missing data—particularly concerning the target variable—and the actigraphy time-series component of the dataset. Our primary objective will be to design and train a deep neural network to effectively address these issues.

1 Background

Problematic Internet Use (PIU) has become a global phenomenon in recent years, with prevalence estimates ranging between 20% and 45% [7]. PIU is defined as excessive and compulsive internet use and is associated with increased symptoms of depression, anxiety, ADHD, and aggression [16, 7, 13], as well as reduced physical activity and life satisfaction [7, 3, 14, 2]. Children and adolescents are particularly vulnerable to PIU, given their developmental stage and the critical impact of early behaviors on long-term health and social outcomes [22, 12].

Measuring PIU in children and adolescents often relies on complex, professional evaluations, creating accessibility challenges due to cultural, linguistic, and resource barriers. In contrast, physical fitness metrics—such as activity levels—are widely available and accessible without clinical expertise. Using these physical indicators (i.e. data gathered with accelerometers) as proxies to identify PIU offers a practical alternative where direct assessments are not feasible. Therefore, predicting PIU from physical indicators helps to establish an early-warning system to inform health and education policies with evidence that can guide resource allocation efficiently.

2 Definitions

Given the challenge, our goal is to find the optimal prediction model for our target variable, the Severity Impairment Index (SII)—a standard measure of problematic internet use. The optimal model is defined as the one that maximizes the quadratic weighted kappa (QWK), and thereby the loss function we will focus on is:

$$\kappa = 1 - \frac{\sum_{i,j} w_{ij} O_{ij}}{\sum_{i,j} w_{ij} E_{ij}} \tag{1}$$

*These authors contributed equally to this work.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

where O_{ij} is the observed ratings matrix, E_{ij} is the expected ratings matrix, and w_{ij} is the weight matrix defined as:

$$w_{ij} = \frac{(i-j)^2}{(N-1)^2} \tag{2}$$

with N being the number of possible rating categories. The expected matrix E_{ij} is calculated as:

$$E_{ij} = \frac{(\sum_k O_{ik})(\sum_k O_{kj})}{n} \tag{3}$$

where n is the total number of rated items.

3 Related Work

Conducting a review of the top Kaggle submissions, we see that the most common approaches are tree-based ensemble learning approaches. Specifically, the top submission with public code uses an ensemble of LightGBM, XGBoost, and CatBoost through a Voting Regressor. This approach achieves a QWK of 0.494, third place overall. As the two submissions above it achieve only marginally better performance of 0.495 and 0.497, it is likely that they employ a similar approach. Almost all the top performing models use a similar architecture, motivating us to begin with a tree-based ensemble approach as a reference, which we can hopefully outperform with our deep neural network.

More generally, there are some key considerations which must be addressed for fitting any model on this dataset. The accelerometer data has significant missingness and is absent entirely for some participants. Furthermore, data for some participants does not differentiate between non-wear and sedentary activity. Previous research has used multiple imputation through chained equations (MICE) and other regression approaches, tree-based models, pooling models, and autoencoders to impute missing actigraphy data [1, 5, 11, 19, 20]. With sufficient data per participant (5+ days of continuous wear or 7+ days with 10+ hours of data each), imputing missing values using simple statistical methods (e.g., mean, median) has been shown to have non-catastrophic effects on results of other psychiatric research [5, 20]. Differentiating between non-wear and sedentary activity is challenging, but existing research using heuristics, logistic regression, and tree-based models has been successful, albeit on larger datasets [18, 19]. For children, separating out school days from non-school days is recommended [5]. The label is missing for a portion of the training data. Other researchers have had success applying pseudo-labelling and other semi-supervised methods to time series with partially labelled training data in related health domains[15, 4, 9].

4 Proposed Method

The Kaggle challenge provides the dataset. The training data consists of 3800 samples, each representing a 5 to 22 year-old. For each sample, we have a row of tabular data on health and demographic characteristics, as well as time series data from the accelerometer. The challenge does allow for the use of other open-source data as external sources, as the data is anonymised, however, matching may be a challenge, so we have not actively considered external datasets yet. We will use a semi-supervised learning (i.e., pseudo-labelling) approach to handle the missing labels[15, 9, 4] and try regression, tree-based, and deep learning methods to impute gaps in the actigraphy time series data[1, 5, 11, 19, 20, 18, 17].

We propose to fit a deep neural network as our main classifier for this problem, in line withe recent literature on the strong empirical results of deep learning applied to time series classification [10]. In particular, we would fit a Hybrid Neural network incorporating both the time series and tabular data, an approach shown to be effective with sensor data [21]. For the time series data, we plan on considering both Long Term Short Memory (LSTM) and Gated Recurrent Unit (GRU) architectures, with an emphasis on the GRU architecture due to its strong empirical performance on smaller datasets with short sequences[17, 8, 6].

References

[1] AE LEE, J., AND GILL, J. Missing value imputation for physical activity data measured by accelerometer. *Statistical Methods in Medical Research 27*, 2 (2018), 490–506.

[2] AL-AMRI, A., ABDULAZIZ, S., BASHIR, S., AHSAN, M., AND ABUALAIT, T. Effects of smartphone addiction on cognitive function and physical activity in middle-school children: a cross-sectional study. *Frontiers in Psychology 14* (2023), 1182749.

[3] ALMAQHAWI, A., AND ALBARQI, M. The effects of technology use on children's physical activity: A cross-sectional study in the Eastern province of Saudi Arabia. *Journal of Medicine and Life 15*, 10 (Oct. 2022), 1240–1245.

[4] BAE, M., SHIN, Y., NAM, Y., LEE, Y. S., AND LEE, J.-G. Semi-Supervised Learning for Time Series Collected at a Low Sampling Rate. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (Barcelona Spain, Aug. 2024), ACM, pp. 59–70.

[5] BORGHESE, M. M., BORGUNDVAAG, E., MCISAAC, M. A., AND JANSSEN, I. Imputing accelerometer nonwear time in children influences estimates of sedentary time and its associations with cardiometabolic risk. *International Journal of Behavioral Nutrition and Physical Activity 16*, 1 (2019), 7.

[6] CAHUANTZI, R., CHEN, X., AND GÜTTEL, S. A comparison of lstm and gru networks for learning symbolic sequences. In *Science and Information Conference* (2023), Springer, pp. 771–785.

[7] CAI, Z., MAO, P., WANG, Z., WANG, D., HE, J., AND FAN, X. Associations Between Problematic Internet Use and Mental Health Outcomes of Students: A Meta-analytic Review. *Adolescent Research Review 8*, 1 (Mar. 2023), 45–62.

[8] CHUNG, J., GULCEHRE, C., CHO, K., AND BENGIO, Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555* (2014).

[9] HEREMANS, E. R. M., PHAN, H., BORZÉE, P., BUYSE, B., TESTELMANS, D., AND DE VOS, M. From unsupervised to semi-supervised adversarial domain adaptation in electroencephalography-based sleep staging. *Journal of Neural Engineering 19*, 3 (June 2022), 036044.

[10] ISMAIL FAWAZ, H., FORESTIER, G., WEBER, J., IDOUMGHAR, L., AND MULLER, P.-A. Deep learning for time series classification: a review. *Data Mining and Knowledge Discovery 33*, 4 (July 2019), 917–963.

[11] JANG, J.-H., CHOI, J., ROH, H. W., SON, S. J., HONG, C. H., KIM, E. Y., KIM, T. Y., AND YOON, D. Deep Learning Approach for Imputation of Missing Values in Actigraphy Data: Algorithm Development Study. *JMIR mHealth and uHealth 8*, 7 (2020), e16113.

[12] LAKKUNARAJAH, S., ADAMS, K., PAN, A. Y., LIEGL, M., AND SADHIR, M. A trying time: Problematic internet use (piu) and its association with depression and anxiety during the covid-19 pandemic. *Child and Adolescent Psychiatry and Mental Health 16*, 1 (2022), 49.

[13] LI, S., WU, Z., ZHANG, Y., XU, M., WANG, X., AND MA, X. Internet gaming disorder and aggression: A meta-analysis of teenagers and young adults. *Frontiers in Public Health 11* (2023), 1111889.

[14] LIU, J., RIESCH, S., TIEN, J., LIPMAN, T., PINTO-MARTIN, J., AND O'SULLIVAN, A. Screen Media Overuse and Associated Physical, Cognitive, and Emotional/Behavioral Outcomes in Children and Adolescents: An Integrative Review. *Journal of Pediatric Health Care 36*, 2 (Mar. 2022), 99–109.

[15] MAMMEN, P. M., ZAKARIA, C., AND SHENOY, P. Personalized Sleep Monitoring Using Smartphones and Semi-supervised Learning. In *Pervasive Computing Technologies for Healthcare* (Cham, 2024), D. Salvi, P. Van Gorp, and S. A. Shah, Eds., vol. 572, Springer Nature Switzerland, pp. 322–338.

[16] RESTREPO, A., SCHEININGER, T., CLUCAS, J., ALEXANDER, L., SALUM, G. A., GEORGIADES, K., PAKSARIAN, D., MERIKANGAS, K. R., AND MILHAM, M. P. Problematic internet use in children and adolescents: associations with psychiatric disorders and impairment. *BMC Psychiatry 20* (2020), 1–11.

[17] SHEWALKAR, A., NYAVANANDI, D., AND LUDWIG, S. A. Performance evaluation of deep neural networks applied to speech recognition: Rnn, lstm and gru. *Journal of Artificial Intelligence and Soft Computing Research 9*, 4 (2019), 235–245.

[18] SKOVGAARD, E. L., ROSWALL, M. A., PEDERSEN, N. H., LARSEN, K. T., GRØNTVED, A., AND BRØND, J. C. Generalizability and performance of methods to detect non-wear with free-living accelerometer recordings. *Nature Scientific Reports 13*, 1 (2023), 2496.

[19] SLYEPCHENKO, A., UHER, R., HO, K., HASSEL, S., MATTHEWS, C., LUKUS, P. K., DAROS, A. R., MINARIK, A., PLACENZA, F., LI, Q. S., ROTZINGER, S., PARIKH, S. V., FOSTER, J. A., TURECKI, G., MÜLLER, D. J., TAYLOR, V. H., QUILTY, L. C., MILEV, R., SOARES, C. N., KENNEDY, S. H., LAM, R. W., AND FREY, B. N. A standardized workflow for long-term longitudinal actigraphy data processing using one year of continuous actigraphy from the CAN-BIND Wellness Monitoring Study. *Nature Scientific Reports 13*, 1 (2023), 15300.

[20] WEED, L., LOK, R., CHAWRA, D., AND ZEITZER, J. The Impact of Missing Data and Imputation Methods on the Analysis of 24-Hour Activity Patterns. *Clocks & Sleep 4*, 4 (2022), 497–507.

[21] YAO, S., HU, S., ZHAO, Y., ZHANG, A., AND ABDELZAHER, T. Deepsense: A unified deep learning framework for time-series mobile sensing data processing. In *Proceedings of the 26th international conference on world wide web* (2017), pp. 351–360.

[22] YU, Y., WU, Y., CHEN, P., MIN, H., AND SUN, X. Associations between personality and problematic internet use among chinese adolescents and young adults: A network analysis. *Journal of Affective Disorders 365* (2024), 501–508.