Toward Healthier Habits: Predicting Problematic Internet Use Through Physical Activity

Zachary Alexander Decker
zad25@cornell.edu

Eytan Rozenblum
er526@cornell.edu

Rachel Minkowitz
rhm256@cornell.edu

Abstract

Problematic internet usage (PIU) among children and adolescents is increasingly recognized as a pressing public health issue, correlated with mental health problems including depression, anxiety, and social withdrawal. Early identification of PIU can facilitate timely interventions and healthier digital habits. This project explores whether easily accessible physical activity and fitness measures can serve as proxies to predict the severity of problematic internet use in children and adolescents. Using data from the Child Mind Institute's Healthy Brain Network (HBN), we integrate demographic information, clinical assessments, physical fitness metrics, and actigraphy-based movement data. After substantial data preprocessing, feature engineering, and dimensionality reduction of the time-series data using an autoencoder, we train and evaluate multiple predictive models. Our model, a LightGBM regressor, achieves a cross-validated quadratic weighted kappa (QWK) of approximately 0.52. These results suggest a moderate relationship between physical measures and PIU severity, while also highlighting the complexity of this predictive task and room for methodological refinements.

1. Introduction

As Gen-Zers, we've seen how technology has shaped our generation - for better and worse. Watching the younger generation struggle with digital addiction has inspired us to take action. This final project comes from a deeply personal place, we're motivated to improve the assessment tools available to address problematic internet usage amongst children today.

1.1. Motivation

With the rapid advancement of technology and its growing influence on society, children today are increasingly at risk of developing problematic internet usage (PIU), often referred to as "Internet Addiction Disorder." Problematic internet use is the over-importance of the internet in ones life to the point of physical, mental and social issues, including depression and anxiety - as shown in figure 1 [2]. As Liu, Jianghong et al. highlight, "Given the adverse public health consequences linked with screen media overuse, understanding these associations is critical for the development of evidence-based guidelines on creating a safe and healthy digital environment for children and adolescents" [4]. While official methods for measuring PIU in young children exist such as the 'Parent-Child Internet Addiction Test', these tools are often complex, inaccessible to the general public or not very robust. On top of that, research shows that avid tech users often display certain physical habits like overweight-ness and reduced physical activity [6].

Our project aims to predict PIU levels in children using physical activity data, which is readily accessible and requires minimal expertise to collect. By identifying patterns associated with PIU, we can potentially enable early interventions, encouraging healthier digital habits and ultimately improving the mental and physical health of the younger generation. Practically, this approach can empower parents, educators, and pediatricians by providing an accessible means to monitor children's digital habits, raising awareness of potential problems earlier and allowing for timely actions like limiting screen time or promoting healthier behaviors. As internet addiction disorder increasingly emerges as a public health concern, this tool could play a meaningful role in addressing the issue and contributing to a healthier digital future [5].

This project and dataset were sourced from a Kaggle Competition where we submitted our models to the leaderboard to evaluate and compare our performance scores.

1.2. Dataset

We utilized the Healthy Brain Network (HBN) dataset, which comprises a clinical sample of about 3,960 participants aged 5–22 years who have undergone extensive clinical and research screenings. This dataset provides a rich array of measures spanning demographic, behavioral, physiological, and clinical domains. Demographic features such as age, sex, and enrollment season offer contextual insights,

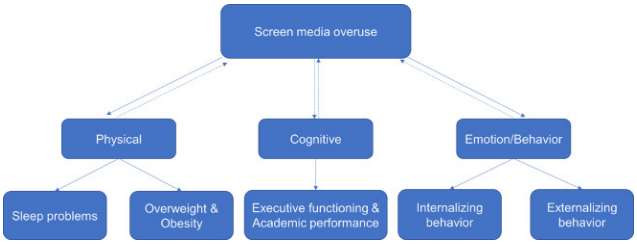

Figure 1. Screen media overuse and associated outcomes.

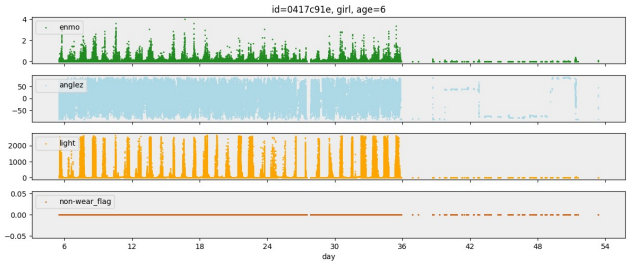

Figure 2. Example of Actigraphy data.

while physical measures like height, weight, BMI, heart rate, and blood pressure capture health and anthropometric profiles. The dataset also includes advanced physical fitness assessments from FitnessGram, which evaluate cardiovascular endurance, muscular strength, and flexibility, alongside detailed body composition metrics derived from bioelectrical impedance analysis (BIA), such as fat mass, lean mass, and body water content.

Self-reported physical activity levels are quantified using the Physical Activity Questionnaires (PAQ-C for children and PAQ-A for adolescents), and sleep quality is assessed through the Sleep Disturbance Scale (SDS), both of which provide additional context relevant to mental health and digital habits. A critical aspect of this dataset is the actigraphy data, consisting of high-frequency accelerometer measurements collected over 30 consecutive days for certain participants, which offer insights into real-world physical activity patterns, sleep-wake cycles, and periods of rest and movement as seen in figure 2. More on this actigraphy data is discussed in section 3.4. Furthermore, the dataset includes responses to the Parent-Child Internet Addiction Test (PCIAT), a 20-item scale measuring behaviors associated with compulsive internet use. The primary outcome variable, the Severity Impairment Index (`sii`), derived from responses to the PCIAT survey, classifies participants into four levels of severity: None, Mild, Moderate, and Severe (1,2,3,4). Although the PCIAT data informs the `sii` score, the true target for this project is the `sii` score. The outcome variable (`sii`) is available for a subset of participants in the training data, but the test set and all its variables remain hidden through the Kaggle competition. Model performance is evaluated using the quadratic weighted kappa (QWK) metric, which accounts for the ordinal nature of `sii` categories and appropriately penalizes misclassifications. QWK basically measures agreement between outcomes while accounting for the ordered nature of the categories. It ranges from 0 (random agreement) to 1 (perfect agreement), with values below 0 indicating worse-than-chance performance. This makes it ideal for scoring predictions where ordinal relationships between categories are essential like `sii` scores.

The quadratic weighted kappa (QWK) is calculated using two matrices.

- The weight matrix $\mathbf{W} \in \mathcal{M}_n(R)$, defined using the squared difference between actual and predicted values.

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

- The histogram matrix $\mathbf{E} \in \mathcal{M}_n(R)$, obtained as the outer product of the actual histogram vector of outcomes and the predicted histogram vector, normalized such that $\mathbf{E}$ and $\mathbf{O}$ (the observed matrix) have the same sum.

Given these matrices, the quadratic weighted kappa is defined as:

$$\kappa = 1 - \frac{\sum_{i,j} W_{i,j} O_{i,j}}{\sum_{i,j} W_{i,j} E_{i,j}}, \tag{2}$$

where $\mathbf{O}$ is the observed $N \times N$ matrix of outcomes.

2. Related Work

As problematic internet usage (PIU) becomes increasingly prevalent, fairly recent studies have explored various approaches to predict variables associated with its occurrence among adolescents, and young adults. For example, in 2023 Geng et al compared the effectiveness of neural network models and linear mixed models in predicting PIU trajectories among adolescents, highlighting the influence of variables such as the familial environment. Using longitudinal data, they analyzed factors like students' self-esteem, screen time, academic performance, and family-related metrics such as parental involvement, family cohesion, and communication patterns. The study found that neural network models were more effective in capturing complex interactions among these variables than traditional linear mixed models [1].

Another study by Jovnic et al. focusing on adolescents between 16-17 years old, examined the impact of sociodemographic factors, internet usage intensity, types of content accessed online, life habits, and affective temperament on PIU. Their findings showed that the Random Forest classifier achieved the best performance, with key predictive variables being time spent on Facebook and cyclothymic temperament. They also used binary classification models like

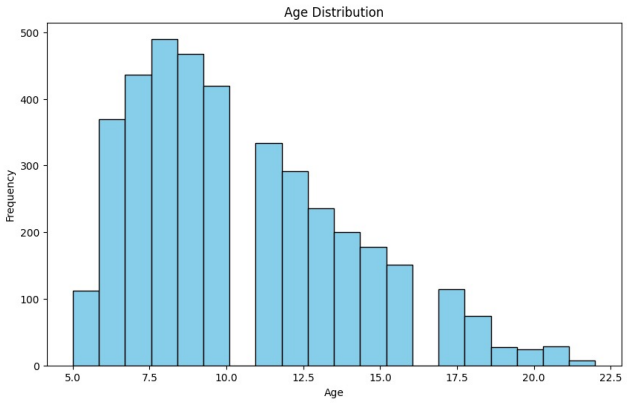

Figure 3. Age Distribution within the HBN Dataset

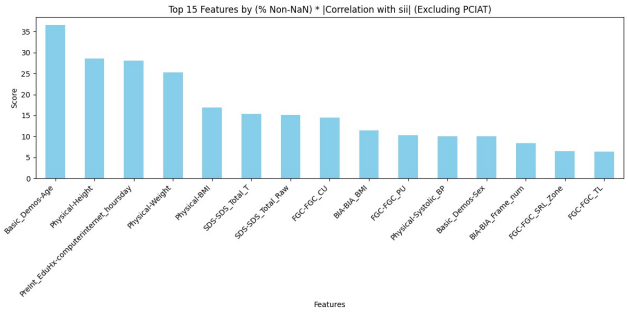

Figure 4. Correlation of features with target variable

Lasso and ElasticNet, and identified cyclothymic temperament, prolonged internet use, and a desire for increased usage as the most critical predictors of PIU when using these methods [3].

Similarly, a study conducted by Tian et al. evaluated the efficacy of neural network models versus linear mixed models in processing psychological longitudinal data from junior high school students. They found that the neural network model exhibited significantly smaller errors compared to the linear mixed model making it more precise and better for prediction [7].

A fourth study conducted recently used a deep learning approach to predict internet addiction among college students. The data they used included survey data about internet addiction, personality, psychological traits, behavioral issues, and social support. They used a 1D Convolutional Neural Network which produced 92.77% accuracy for distinguishing between normal internet users and those who were addicted. Their study concluded that factors such as being a second-year student, having depression or anxiety and having overly controlling families were significant contributors to internet addiction [8].

Our final project focuses on the population of children, who tend to be underrepresented in PIU prediction studies (Figure 3). A key distinction of our work is the use of wearable device data in conjunction with demographic and survey data to predict the likelihood of a child's Problematic internet usage. Our goal is to create a practical method for predicting and addressing PIU in this younger age group, ultimately contributing to more awareness and eventually healthier digital habits and well-being.

3. Methods

3.1. Exploratory Data Analysis

We performed EDA to understand the characteristics and completeness of the dataset. We began by examining the distribution of missing values, which varied significantly across different features. Many measures were sparsely recorded, prompting us to take into account the 80%-90% of missingness in some categories through careful imputation. This initial step guided our feature selection process, ensuring we concentrated on a subset of informative variables that were available for a sufficient fraction of participants.

Then we performed some demographic analysis on age, sex, and trial enrollment season. We found that sex and season were fairly balanced with either being strongly correlated with other features. Season was by far the most unimportant and was dropped from further analysis. Sex was slightly imbalanced and showed some correlation especially with older participants. But since most participants were pre-adolescents, we considered sex an unimportant feature. This is a large simplifying assumption, and may merit another look to improve predictive accuracy. Age was highly correlated with most other features, leading us to engineer interaction terms as described in the feature engineering section.

We also investigated the relationships between features and the target variable, `sii` as seen in figure 4. We calculated correlation coefficients between each candidate predictor and `sii`, identifying modest but consistent relationships for certain demographic and physical measures. Age, for example, showed a moderate positive correlation with `sii`, suggesting that older adolescents might exhibit different internet usage patterns than younger children. Similarly, certain body composition metrics, as well as reported daily internet usage hours, displayed non-negligible correlations with `sii`. By sorting features by their absolute correlation or by the product of their correlation and data completeness, we obtained prioritized lists that helped us focus on a core set of predictors.

Feature importance analysis performed later in the modeling phase - particularly with tree-based models like LightGBM - reinforced these EDA findings. Variables related to demographics (e.g., age and sex), body composition, and self-reported internet hours consistently ranked highly, indicating that these dimensions contained discernible signals about problematic internet usage severity as seen in figure

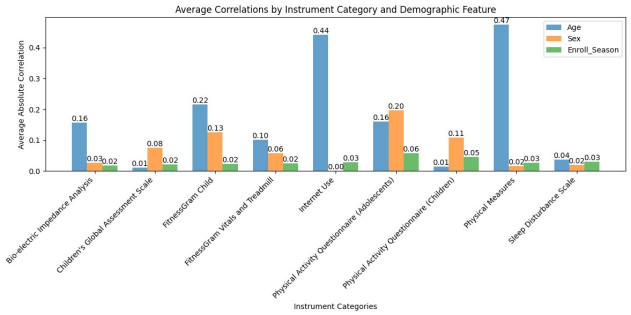

Figure 5. Age proves to be highly correlated to certain features - likely changes interpretation for certain features.

5. Actigraphy-based features, after being compressed into lower-dimensional embeddings via an autoencoder, also contributed slightly to predictive performance. Although these embeddings did not produce strong correlations individually, their latent structure captured subtle patterns of activity and rest that were not fully encapsulated by traditional summary statistics. This underscores the value of using representation learning methods to leverage complex datasets.

Overall, the EDA provided critical insights into the nature of the data and the predictive potential of various features. By combining correlation analyses, completeness metrics, and feature importance estimates from preliminary models, we were able to streamline our modeling strategy. This foundational work helped us discard uninformative variables, focus on key predictors, and ultimately develop a more adapted model.

3.2. Benchmarking

Initial modeling efforts employed OLS linear regression and random forest to establish a baseline. We used three models: (i) using all the features in the dataset, (ii) using top 15 features (with features ranked in decreasing order of correlation with `sii`, weighted by the percentage of non-NaN elements for a feature, $1 - \%_{NaN}$), and (iii) using top 15 features ranked by Gini importance with random forest. Results are reported in Table 1.

Features	OLS	Random Forest
All HBN Features	0.349	**0.390**
Top15 (Correlation)	**0.354**	0.345
Top15 (Gini)	0.338	0.351

Table 1. Baseline QWK across different feature selection strategies

In the one page milestone submission, we reported much higher baseline QWK scores around 0.9. This is because we were using the PCIAT features which were really unavailable in the complete test set when accessed through the Kaggle competition. The use of the PCIAT to predict the `sii` score made the problem much simpler (more discussion about this in section 3.4).

Due to how Quadratic Weighted Kappa is calculated, random guessing would result in a score very close to zero. All models achieved scores above .33, indicating significant progress needs to be made but the problem is tractable. The best performing baseline model was Random Forest using all the features followed by OLS using the correlation picked features. Random forest may perform better even with the less helpful features due to its robustness to missing data and varying datatype, but with such small differences and other confounding variables, more testing is required. The features chose by Gini and by our own naive correlation metric were similar, explaining the small variation in accuracy results.

3.3. Data Preprocessing

Given the complexity of the dataset and the substantial proportion of missing values, careful preprocessing was essential. Initially, we quantified missingness for each feature, prioritizing those with higher completeness to ensure reliable predictive modeling. Once these core features were identified, KNN-based imputation was applied to fill gaps, chosen for its ability to preserve local structures in the data distribution. Other imputation strategies, such as median filling and imputation with Random Forrest estimators, were tested but provided no consistent benefit or were too computationally expensive for the amount of data we're working with.

Categorical variables, including demographic attributes and various discrete classification schemes (e.g., fitness zones), were encoded to ensure compatibility with the machine learning frameworks. Although outliers were present in some measures, we retained them to preserve potentially informative variance, given the complexity of the health indicators and the possibility that unusual values might reflect genuinely extreme cases rather than noise.

3.4. Feature Selection and Engineering

Feature selection proceeded with a dual emphasis on both data availability and predictive potential. We focused on variables that were frequently recorded and demonstrated at least a modest correlation with `sii`. Key predictors included age, BMI, and reported daily hours of internet use. To enhance the model's capacity to capture complex interactions, we introduced engineered features. The choice to engineer certain features were as a result of the strong demographic correlations shown in figure 5 above. Those features include the 'Activity_Level_Age' feature which represents the product of a participant's age and the results from the bio-electric Impedance Analysis activity level (very light, light, moderate, heavy, exceptional), as well the 'GSND_Zone_Age' feature which represents the product of a participant's age and their grip strength (strong,

normal, weak). Our engineered features anticipate that older children/adolescents might exhibit distinct relationships between physical indicators and internet usage severity. This intermediate step enabled the model to uncover nonlinear and context-dependent patterns rather than relying solely on raw measurements.

Notably, PCIAT features were excluded for training because `sii` is derived directly from the PCIAT responses and PCIAT data was unavailable in the test set as we are trying to use other sources of information for this exact prediction.

3.5. Actigraphy Data Representation

One of the most challenging components was the actigraphy data, a high-volume time series reflecting continuous wrist-worn accelerometer readings over multiple days. Directly modeling the raw time series was computationally infeasible and prone to overfitting. Additionally it is difficult to create a model that can effectively utilize both time-series data and tabular data. To tackle this, we first calculated a suite of statistical summaries - means, medians, standard deviations, minima, and maxima- for each participant's activity streams. These summaries captured essential activity patterns while reducing data dimensionality.

To further refine representation, we employed an autoencoder, a neural network trained to compress the actigraphy-derived feature space into a lower-dimensional latent embedding. Through this encoder-decoder structure, the model learned to retain the core aspects of a participant's activity profile. These compact, autoencoder-derived embeddings were then integrated back into the main tabular dataset, effectively enriching each participant's feature vector with a learned representation of their daily activity dynamics.

3.6. Final Model

From our data analysis and research, we concluded on using a Gradient Boosting model. Similar to Random Forest, Gradient Boosting builds an ensemble of decision trees, but instead of training them independently, it trains each new tree to correct the errors made by the previous ones. This sequential learning process allows Gradient Boosting to achieve higher accuracy by focusing on the hardest-to-predict examples, improving the model's performance over time. Gradient Boosting is robust to lossy data because it can handle missing values by building trees that adapt to the available information. It also works well with different data types, such as categorical and numerical features, by efficiently splitting on both types of variables without requiring complex preprocessing. Ultimately, gradient boosting - specifically using LightGBM (Light Gradient Boosting Machine) - offered the best balance of efficiency, interpretability, and predictive accuracy. LightGBM's ability to handle a mix of numeric and categorical features and to efficiently search over a high-dimensional feature space quickly proved advantageous. Through iterative tuning of hyperparameters, such as the number of leaves (= 31), learning rate (= 0.05), and feature sampling fractions (= 0.9), and by employing 5-fold cross-validation for robust performance estimation, we arrived at a model configuration that outperformed earlier attempts. Although we cast the prediction task as a regression problem (targeting continuous `sii` scores), the model's final outputs were rounded to the nearest integer category before evaluation with QWK, aligning them with the ordinal nature of the `sii` scale.

4. Results

The final LightGBM model, which incorporated most of the non-PCIAT HBN tabular features and specially engineered features, autoencoder-derived actigraphy embeddings, and KNN-imputed missing values, achieved a cross-validated QWK of approximately **0.52** (with a standard deviation around 0.03) and an RMSE of roughly 0.618 (±0.0285). Additionally we scored 0.33 on the Kaggle Competition Leaderboard based on the full test set - while the highest score on the Leaderboard is currently 0.502. This suggests that the training data doesn't generalize very well to the hidden test dataset. Meanwhile, these results represent a substantial improvement over the baseline linear and random forest models, demonstrating that integrating diverse sets of physical and behavioral measures can yield meaningful, if moderate, predictive power for problematic internet use severity.

Examining feature importance and correlation patterns reveals that age, body composition measures (notably BMI and related anthropometric indicators), and self-reported daily internet hours emerged as critical variables. While actigraphy embeddings individually exhibited only moderate correlation with `sii`, their collective latent representation provided a small yet consistent boost in performance. The improvement associated with the embeddings suggests that subtle daily activity patterns—captured but not explicitly recognized by manual feature engineering—hold some predictive value.

5. Discussion

Our results show that physical indicators can reflect certain patterns, but predicting problematic internet usage or its severity index accurately likely take deeper psychological and environmental factors into account. Here are some things that stood out, where we struggled, and possible future work that could improve our outcomes.

5.1. Key Findings

Our findings indicate that physical activity, body composition, and demographic variables offer a modest but non-

trivial degree of predictive insight into problematic internet usage severity. Achieving a QWK near 0.52 suggests that these measures capture underlying behavioral or health-related signals correlated with internet use patterns. This level of agreement, however, remains moderate, implying that while physical indicators can partially reflect PIU risk, a substantial portion of the variance may lie in psychosocial or environmental factors not captured by the available features.

5.2. Limitations

Despite carefully structured preprocessing and modeling, the predictive performance was constrained by data sparsity and the inherent complexity of PIU. Data types and sources varied throughout the dataset leading to high levels of missingness for certain features. Ultimately the dataset required extensive imputation, likely diminishing the richness of the information. Moreover, while the actigraphy data is voluminous, it proved challenging to fully exploit: without more advanced temporal modeling approaches, the simplified embeddings may have omitted crucial temporal structures relevant to PIU. Additionally, treating `sii` as a numeric target and then rounding predictions may have introduced distortions, as the `sii` categories are ordinal rather than strictly continuous.

5.3. Future Work

To start, once the Kaggle competition is complete (December 19th, 2024) and we have access to the full test set, we will be able to complete more thorough testing when building our models.

Moving forward, research could focus on more sophisticated temporal modeling strategies, employing sequence models or transformers to capture diurnal patterns, weekend vs. weekday variability, or sleep/wake cycles hidden in the actigraphy data. The poor importance of the actigraphy derived features in the final model may suggest the auto-encoding approach was sub-optimal, potentially destroying or obfuscating important data. Researching further into ways to interpret lossy time-series data could prove useful.

Implementing ordinal regression methods or custom loss functions aligned with QWK might also yield better calibration of predictions. Moreover, integrating psychosocial indicators - such as family environment, mental health assessments, or subjective well-being measures - could bridge the gap between purely physical signals and the complex psychological underpinnings of problematic internet use. By enriching both the breadth and depth of the data and refining the modeling framework, we may progress toward more accurate and meaningful predictions that support early interventions and healthier digital habits.

Disclaimer: ChatGPT was used for python troubleshooting and semantic improvement for this report.

References

[1] X. Geng, J. Zhang, Y. Liu, L. Xu, Y. Han, M. N. Potenza, and J. Zhang. Problematic use of the internet among adolescents: A four-wave longitudinal study of trajectories, predictors and outcomes. *Journal of Behavioral Addictions*, 12(2):435–447, Jun 2023.

[2] I. Goldberg. Internet addiction disorder. *CyberPsychology & Behavior*, 3(4):403–412, 1996.

[3] J. Jović, A. Ćorac, A. Stanimirović, M. Nikolić, M. Stojanović, Z. Bukumirić, and D. Ignjatović Ristić. Using machine learning algorithms and techniques for defining the impact of affective temperament types, content search and activities on the internet on the development of problematic internet use in adolescents' population. *Frontiers in Public Health*, 12, 2024.

[4] J. Liu et al. 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):99–109, 2022.

[5] S. K. Muppalla, S. Vuppalapati, A. R. Pulliahgaru, and H. Sreenivasulu. Effects of excessive screen time on child development: An updated review and strategies for management. *Cureus*, 15(6):e40608, Jun 2023.

[6] J. M. Nagata, N. Smith, S. Alsamman, C. M. Lee, E. E. Dooley, O. Kiss, K. T. Ganson, D. Wing, F. C. Baker, and K. P. Gabriel. Association of physical activity and screen time with body mass index among us adolescents. *JAMA Network Open*, 6(2):e2255466, Feb 2023.

[7] M. Tian, Q. Xing, X. Wang, X. Yuan, X. Cheng, Y. Ming, K. Yin, Z. Li, and P. Wang. Prediction of junior high school students' problematic internet use: The comparison of neural network models and linear mixed models in longitudinal study. *Psychology Research and Behavior Management*, 17:1191–1203, Mar 2024.

[8] X. Wang, E. Zhang, Y. Cui, J. Huang, and M. Cheng. Predicting internet addiction in college students using a 1d-cnn model: Analysis of influencing factors. *Dyna*, 91(233):66–74, Aug 2024.