Data Exploration

Predicting Problematic Internet Use in Children Utilizing Internet Usage Behavior Data & Physical Activity Data

Sujen Kancherla
skancherla@ucsd.edu
Juo-Hsuan Chang
juc077@ucsd.edu

1 Introduction

The topic we chose is an ongoing Kaggle competition aimed at detecting Problematic Internet Use (PIU) among children. The increasing use of the Internet among children has been a growing topic of discussion in recent years, especially with social media and the unlimited online content that is fed to everyone. While the Internet offers numerous benefits, excessive use has been linked to mental health issues such as depression, anxiety, and social isolation. Current methods for identifying PIU at an early stage are often complex and require professional involvement. In contrast, physiological signals captured through wearable devices have become widely obtainable, offering an alternative for evaluating PIU.

In this project, we utilized the Child Mind Institute - Problematic Internet Use dataset [1] to investigate the potential of predicting PIU using internet usage behavior data and physical activity data. Our main objective is to predict the Severity Impairment Index (SII), which reflects a child's mental health, based on their internet use, physical activity, sleep patterns, and other related factors in the dataset. This involves two key steps: (1) data exploration to understand behavioral patterns and (2) development of a predictive model to assess mental health conditions based on the data.

The code for this project is available on GitHub: https://github.com/UCSDNaNaNa/Project_225A.git

2 Data Exploration

The dataset used in this project is sourced from Kaggle: Child Mind Institute - Problematic Internet Use [1]. The dataset contains training and testing subsets. Our analysis primarily focused on the training dataset, which includes data from 3,960 subjects. After filtering out the entries with abnormal or missing values, 2,736 subjects remained. Furthermore, wrist-worn accelerometer data is available for only 996 subjects.

2.1 Internet Usage Behavior Data

The Internet Usage Behavior Data includes demographic and physical attributes, such as age, sex, height, weight, and body mass index (BMI), along with fitness-related metrics like grip strength, push-up count, and trunk lift value. We visualized the distribution of these features using histogram (Figure 1) to identify patterns.

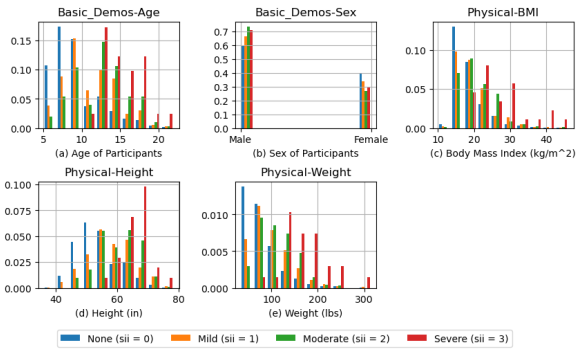

Figure 1: Histogram Visualization

From Figure 1, participants with *sii = 3* (Severe) tend to be older with higher BMI, height, and weight. These trends are further confirmed through calculation of the mean values shown in Figure 2.

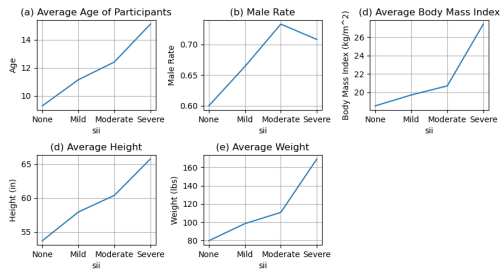

Figure 2: Average Value of Features

2.2 Wrist-worn Physiological Data

The Wrist-worn Physiological Data are data collected from some participants wearing an accelerometer for up to 30 days continually. The data includes acceleration on the X, Y, and Z axes in the gravity unit, and derived metrics such as the Euclidean Norm Minus One value (ENMO):

$$enmo[i] = |\sqrt{X^2[i] + Y^2[i] + Z^2[i]} - 1|. \quad (1)$$

Other features include *non-wear_flag*, *anglez*, *light*, *battery_voltage*, *time_of_day*, *weekday*, etc. However, according to the feature *non-wear_flag*, the average wearing ratio is only 22.20%, indicating a significant limitation in physiological data availability.

We extracted statistical features (e.g., mean, max, min, range, median, mode) and the interpretive features specific to ENMO values under certain conditions. These include:

- *enmo_avg*: Average ENMO value from the entire signal.
- *enmo_wear*: Average ENMO value when *non-wear_flag* == 0.
- *enmo_night*: Average ENMO value when *non-wear_flag* == 0 during nighttime (11 pm - 4 am).
- *enmo_day*: Average ENMO value when *non-wear_flag* == 0 during daytime (8 am - 6 pm).
- *enmo_high*: Ratio of the ENMO values exceeding 2.

Figure 3 and Figure 4 present the histogram and mean values of these features, respectively. We grouped the participants with *sii = 2* and *sii = 3* to improve analysis reliability due to limited data. The trend is evident in both Figure 3 and Figure 4. Participants with higher SII present a greater proportion of low ENMO value, indicating low activity levels. In addition, the average values of all ENMO-related features decrease as the SII category increases.

2.3 Severity Impairment Index

Severity Impairment Index (SII) is a standard measure of PIU. The participants' parents were required to answer 20 questions associated with use of the Internet including compulsivity, escapism,

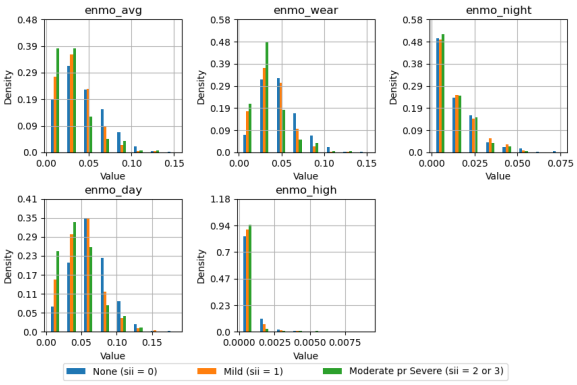

Figure 3: Histogram of ENMO-Related Features

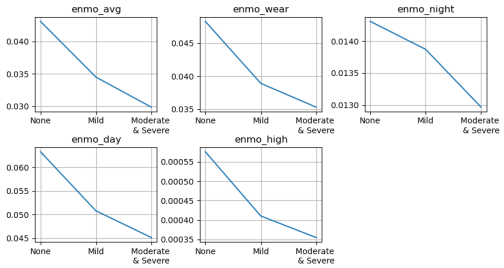

Figure 4: Average Value of ENMO-related Features

and dependency to evaluate their children's behaviors. Based on the responses, children are categorized into 4 levels: None (0), Mild (1), Moderate (2), and Severe (3).

The dataset exhibits significant class imbalnace, as shown in Table 1. For example, only 1.24% of participants in the full dataset fall into the Severe (*sii = 3*) category, and this proportion drops further in subsets containing physiological data. Specifically, only 10 participants with physiological data are classified as *sii = 3*. This imbalance presents challenges for model training and evaluation.

	Severity Impairment Index				
Dataset	Total	0	1	2	3
Entire Dataset	3960	-	-	-	-
Remove SII = *NaN*	2736	1594	730	378	34
Physiological Data	996	583	266	137	10
Merged Data	990	580	264	136	10

Table 1: Data size among different data frames.

3 Preprocessing

3.1 Weird Value Handling

In the preprocessing stage, we handle weird or out-of-range values in the dataset. One notable case

is the body fat values, which should logically fall between 0% and 50%. However, the dataset contains many outliers, including some negative values. To handle this, we replace the out-of-range values with the median body fat value, ensuring that all entries fall within the valid range.

The procedure can be outlined as follows:

1. Identify body fat values that are either less than 0 or greater than 50.
2. Replace these out-of-range values with *NaN* (Not a Number).
3. Impute the missing values by replacing them with the median body fat value in the dataset.

We discover that the irregular values not only appear in body fat, but all the other features obtained by Bio-electric Impedance Analysis (BIA). To address this, we replace all BIA-obtained values with the median of the corresponding feature in the subset of data containing irregular body fat values. This preprocessing step helps clean the data, allowing models to train on more realistic and representative values.

3.2 Feature Selection

Feature selection is a crucial step for improving model performance by choosing the most relevant variables. In this dataset, we start by selecting a subset of features that have a known or suspected relationship with the target variable, *sii* (which indicates problematic internet use).

The selected features include demographic, physiological, and behavioral variables such as:

- Demographics: *Basic_Demos-Age*, *Basic_Demos-Sex*
- Physical characteristics: *Physical-Height*, *Physical-Weight*, *BIA-BIA-BMI*, *BIA-BIA-Fat*, and *BIA-BIA-SMM*
- Fitness measures: *Fitness_Endurance-Max_Stage*, *Fitness_Endurance-Time_Mins*, *Fitness_Endurance-Time_Sec*
- Behavior-related features: *PreInt_EduHx-computerinternet_hoursday*, *FGC-FGC_CU*, *SDS-SDS_Total_Raw*

A correlation analysis is then performed to identify the strength of the relationship between each feature and the target variable. This allows us to retain only the features that have the most significant correlations with *sii*.

We also refine our feature set by removing features that either have little to no correlation with *sii* or are redundant. The feature selection process significantly reduces the dimensionality of the data, focusing on the most predictive attributes.

Here are the correlations that some features have:

Feature	Correlation with SII
Basic_Demos-Age	0.365990
Physical-Height	0.360802
PreInt_EduHx-computerinternet_hoursday	0.331288
Physical-Weight	0.316121
BIA-BIA_LDM	0.261165
BIA-BIA_BMR	0.254064
FGC-FGC_CU	0.200546
BIA-BIA_TBW	0.297511
BIA-BIA_SMM	0.271701
BIA-BIA_Fat	0.183037
CGAS-CGAS_Score	-0.078733
Basic_Demos-Sex	-0.100148
non_wear_ratio	0.144768
enmo_day	-0.259522
enmo_wear	-0.234342
enmo_avg	-0.208179
enmo_std	-0.2081

Table 2: Correlation of features with SII.

4 Inference Model

4.1 Baseline Model

In our baseline models, we implemented a logistic regression using the scikit-learn package on the physiological data with the original training data combined. The main issue with this approach was the dataset being very small. There were many columns missing some data for some rows in the physiological and original training data. To address this, specifically with the out of range Body Fat numbers, we imputed the out of range numbers with the median body fat.

4.2 Improved Models

The next model we implemented was a decision tree. This makes sense intuitively because it can lead to interpretability of the model. The decision tree can help to pose real questions related to the features to any parents that may help them understand better what features relate to the problematic internet use. Here are some examples of decisions the model came up with after training:

- For a certain weight threshold model chooses more problematic for higher weights

- Age is the first decision, making the model lean towards certain classes but still keeps uncertainty until looking at other features

4.3 Model Results

The models evaluated in this study achieved the following accuracies:

Model	Dataset	Feature Selection	Accuracy (%)
Logistic Regression	IUPD	None	60.40
Logistic Regression	WWPD	None	60.50
Logistic Regression	Merge	None	59.60
Logistic Regression	Merge	Yes	62.63
Decision Tree	Merge	Yes	58.08
SVM	Merge	Yes	59.60

Table 3: Model accuracies for predicting problematic internet use using various features and datasets. IUPC: Use only Internet Usage Behavior Data; WWPD: Use only Wrist-worn Physiological Data; Merge: Use both Internet Usage Behavior Data and Wrist-worn Physiological Data.

5 Conclusion

The results show very similar results with the models, where the decision tree was ideally the better as it performed similarly and also has interpretability. This project really illustrates how some of these features can correlate too the problematic internet usage. We can say that it proved our insights before starting. For example, it makes sense that children with less physical activity might be prone to more problematic internet usage, and the models all show this.

5.1 Future Implications

The results of this study have significant implications for understanding problematic internet use. By identifying key predictors, the methodology paves the way for:

- **Targeted Interventions:** The selected features could be used to develop interventions targeting individuals at higher risk of problematic internet use, focusing on both physical and behavioral health.
- **Model Generalization:** Future work could explore the applicability of this approach to other populations or datasets, potentially enhancing the generalizability of the findings.
- **Improved Methodologies:** Further research could focus on advanced feature engineering techniques or deep learning approaches to refine prediction models.
- **Longitudinal Studies:** Incorporating longitudinal data could help establish causality and better understand the temporal dynamics of internet use and its correlates.

References

[1] Adam Santorelli, Arianna Zuanazzi, Michael Leyden, Logan Lawler, Maggie Devkin, Yuki Kotani, and Gregory Kiar. Child mind institute — problematic internet use. https://kaggle.com/competitions/child-mind-institute-problematic-internet-use, 2024. Kaggle.