Deep in the Web: using deep learning methods to predict Problematic Internet Use in today's youth

Evan Matthews[1], Vikram Ramavarapu[1], and Krishnaveni Unnikrishnan[1]

[1]CS 412 Group G6

December 11, 2024

Abstract

The Internet's pervasive role in modern life has raised concerns about Problematic Internet Use (PIU), particularly among children and teens. Our research aims to predict early signs of PIU using machine learning techniques applied to data from the Child Mind Institute's Healthy Brain Network. This study employs a comprehensive methodology combining both cross-sectional and time-series data for future analysis. Initial results from multiple models, including Random Forest, XGBoost, SVM, and Feed Forward Neural Networks, demonstrate promising accuracy rates, with XGBoost achieving the highest mean accuracy of 0.682. Our project experimentation is structured in three phases: data preprocessing, initial model evaluation, and fine-feature reevaluation. The methodology incorporates innovative approaches such as sequential modeling for time-series data and ensemble techniques combining cross-sectional and sequential models. Preliminary findings suggest that machine learning can effectively predict PIU severity using quantitative measures compared to traditional assessments. This research contributes to the growing field of digital health by providing a data-driven approach to identifying at-risk youth for PIU. Code for the analysis can be found at this Github Repository.

1 Introduction

The Internet has become an integral part of our daily lives, with people of all ages spending a significant amount of time online. This trend has given rise to concerns about the potential impacts of excessive internet use, particularly on children and teens. Problematic Internet Use (PIU) is a condition characterized by excessive or poorly controlled preoccupations, urges, or behaviors regarding computer use and internet access that lead to impairment or distress [4]. PIU has been associated with a range of mental health issues, including depression, anxiety, and impulsivity [2]. As such, identifying early signs of PIU in children and teens is crucial for prevention and intervention. Despite having multiple studies showing the negative effects of excessive internet use, exact details about PIU warning signs and the most at-risk individuals are still unknown. These studies can be useful, but they also introduce biases and often fail to show the true factors which correlate a participant's estimated internet impact [1, 5]. In this project, we aim to predict early signs of PIU in children and teens using machine learning techniques, leveraging data from the Child Mind Institute's Healthy Brain Network. The project plan consists of three phases: data preprocessing, initial model evaluation, and fine-feature reevaluation. We will submit our work to the Child Mind Institute's (CMI) Kaggle competition on PIU prediction at a later date.

2 Motivation

With the rise of machine learning and pattern prediction models, the ability to analyze and predict upon more complex data and parameters becomes much more approachable. Likewise, child development is a multi-facted situation in which parenting and environmental factors can lead to an incredibly high number of outcomes. This field has had great strides in classical research, but a more modern approach could lead to significant development in the success of future generations. Additionally, predictions against an extensive number of possible outcomes like this represents a current roadblock in machine learning- that is, how modern predictive models can adapt to an ever-increasing set of parameters and decreasing set of training data. Finally, child psychology is interested in recognizing patterns in early behavior in order to reduce the impact of harmful effects from a child's environment.

Despite having multiple studies showing the negative effects of excessive internet use, exact details about PIU warning signs and the most at-risk individuals are still unknown. These studies can be useful, but their results focus primarily on written or binary feedback from students or parents. Additionally, they introduce biases and often fail to show the true factors which correlate a participant's estimated internet impact [1, 5].

Another major drawback of assessing PIU is in its subjective nature. Problematic internet use is characterized by many different variables that are hard to measure. As such, using quantitative measures such as the Severity Impairment Index (SII) allow for the application of data mining methods to aid in the classification of PIU severity. Moreover, other measurable attributes such as sleep quality and duration, physical activity level, and duration of internet usage can all be used to understand correlations with PIU. This project intends to rectify these issues by using a machine-learning approach to predict early signs of PIU using a wider range of variables on children and teens.

3 Related Work

Research on Problematic Internet Use (PIU) has gained significant attention due to its increasing prevalence and association with various psychological and behavioral issues. Early investigations into PIU highlighted its similarities with substance use disorders, impulse control disorders, and obsessive-compulsive disorder. Studies have revealed concerning prevalence rates between 1.5% and 8.2% in the United States and Europe, emphasizing the growing social impact of this condition [2]. The relationship between PIU and psychiatric disorders has been extensively documented, with research showing significant associations with depressive disorders and attention-deficit/hyperactivity disorder (ADHD). A notable study found that individuals with PIU were more than twice as likely to have depressive disorders ($aOR = 2.43$), and showed increased likelihood of having ADHD combined presentation ($aOR = 1.91$) and Autism Spectrum Disorder ($aOR = 2.24$) [5].

Recent investigations have focused on understanding the personality profiles and emotional factors contributing to PIU. Research has identified specific personality traits associated with PIU, including lower scores in novelty seeking, harm avoidance, and reward dependence. Additionally, emotional dysregulation has emerged as a significant factor, with studies suggesting that PIU may serve as a behavioral mechanism for escaping negative affects. Treatment approaches for PIU have primarily centered on addressing comorbid conditions, with cognitive behavioral therapy and selective serotonin reuptake inhibitors showing promise as potential interventions. However, researchers emphasize that detailed treatment guidelines require further investigation, particularly given interactions between PIU and various psychological disorders.

Currently, the field continues to evolve, and debates haved continued regarding diagnostic criteria and classification. While the Internet's positive impact on well-being is widely acknowledged, the pathological aspects of its use remain understudied, particularly regarding subtle psychological changes such as online disinhibition. This highlights the need for additional research into the pathophysiology, epidemiology, natural course, and treatment of PIU to develop more effective intervention strategies. In terms of our current work, given that the original scope of the project was accepted, we are pressing forward with this plan with no significant changes. The most crucial critique provided- that the validation plan and evaluation metric were not clear- are likewise addressed in the methodology section.

4 Methodology

Data for this project has two components: cross-sectional, and sequential (time-series). The cross-sectional data is per participant and contains fields described in the following table. Each sequential dataset is per participant and each entry of the dataset represents the status of the participant's heartrate monitor at a given point in time. PCIAT is the Parent-Child Internet Addiction Test score, which is used to compute the Severity of Internet Addiction Index (SII) score. The SII score is the target variable for this project. For the description of fields in the time-series dataset, see Table A.

The project is divided into three phases: data preprocessing, initial model evaluation, and fine-feature reevaluation. The data preprocessing phase entails dropping survey-based fields used to compute PCIAT, which is then used to compute the SII, as our model's intention is

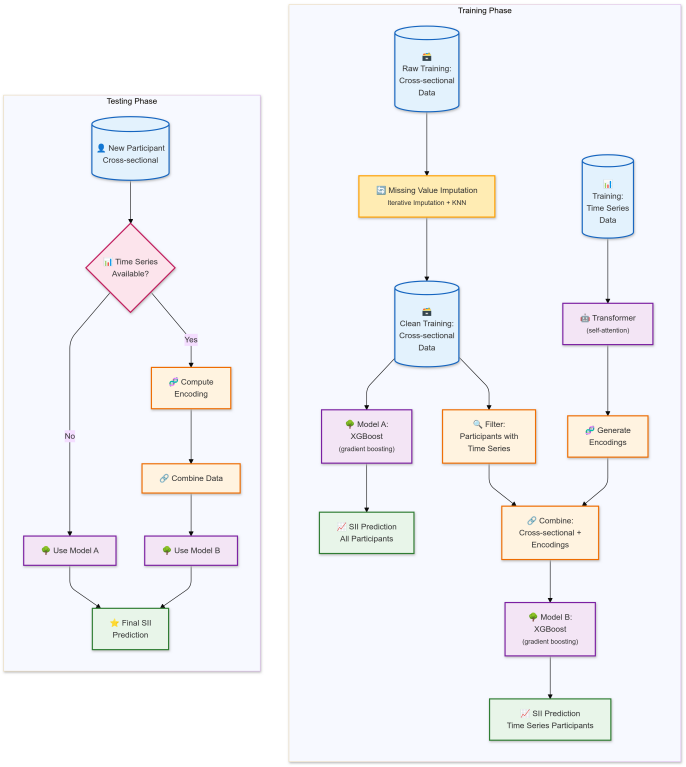

Figure 1: Model architecture

to compute SII directly from the other metrics. Missing values in the data are filled using iterative imputation, and the missing SII values are filled in using K-Nearest Neighbors ($k = 5$).

Multiple models are evaluated on the cross-sectional data: Random Forest, XGBoost, SVM, and a feed forward neural network. After this, a sequential model, evaluated amongst transformers or auto-encoders, is trained on the time-series data. The sequential model allows us to compute an embedding of the time-series data, which will be used as an additional feature in the cross-sectional model. The final model is an ensemble of the cross-sectional and sequential models, with the sequential model's embedding as an additional feature in the cross-sectional model. The classifier model is retrained on the concatenated dataset, to predict the SII. Finally, validation of the trained models is performed using 10-fold cross-validation, with the best model selected based on performance metrics.

After comparing classifiers and selecting the best model, we land on the architecture shown in Figure 1. First, XGBoost is trained on the cross-sectional data, where missing values are filled using label propagation. A transformer encoder is trained on the time-series data using reconstruction loss, and the encoder is used to compute an embedding of the time-series data. The embedding is concatenated with the cross-sectional data, and the XGBoost model is retrained on the concatenated dataset. In the testing phase, for a datapoint that has a time-series component, the transformer encoder is used to compute the embedding, which is then concatenated with the cross-sectional data and fed into the XGBoost model to predict the SII. If time-series data is not available, the XGBoost model is used to predict the SII directly from the cross-sectional data.

4.1 Results

4.1.1 Cross-Sectional Data

Preliminary results of 10-fold cross-validation provide insight into the performance consistency of each model- Random Forest Classifier, XGBoost Classifier, Support Vector Classifier, and Feed Forward Neural Network across different subsets of the dataset. This method helps ensure that the reported accuracy is not overly dependent on any particular training subset, giving a more reliable view of how each model would perform in a real-world setting. These results are summarized in Table 1 and Table 2.

Model	**Fold 1**	**Fold 2**	**Fold 3**	**Fold 4**	**Fold 5**	**Fold 6**	**Fold 7**	**Fold 8**	**Fold 9**	**Fold 10**
RF	0.684	0.682	0.682	0.674	0.687	0.674	0.684	0.707	0.649	0.672
XGB	0.689	0.667	0.669	0.689	0.682	0.684	0.674	0.732	0.636	0.694
SVC	0.649	0.657	0.646	0.649	0.649	0.646	0.649	0.649	0.654	0.652
FFN	0.710	0.649	0.669	0.692	0.689	0.684	0.687	0.674	0.649	0.694

Table 1: 10-Fold Cross-Validation Results for Each Model

Model	**Mean Accuracy**
Random Forest Classifier	0.680
XGBoost Classifier	0.682
Support Vector Classifier	0.650
Feed Forward Neural Network	0.680

Table 2: Mean Accuracy for Each Model

Using hypothesis testing, we conclude that XGBoost model is significantly better than the other models based on a Student's t-test at significance level $\alpha = 5\%$ and number of parameters to be trained.

4.1.2 Sequential Data

The sequential model was trained on the time-series data using a transformer encoder. The model was trained using reconstruction loss, and the encoder was used to compute an embedding of the time-series data. The loss curve for the transformer model is shown in Figure 2.

After 10 epochs of training, each patient with time-series data has an embedding computed, which is then used as an additional feature in the cross-sectional XGBoost model. The XGBoost model with the same maximum depth ($depth = 15$), when trained on the concatenated dataset, achieves a lower validation accuracy than the model trained on the cross-sectional data alone. However, the model achieved 100% accuracy on the training data, suggesting overfitting. As such, the XGBoost model was retrained on the concatenated dataset with a lower maxmimum depth ($depth = 10$). The 10-fold cross-validation results for the retrained model are shown in Figure 4.1.2.

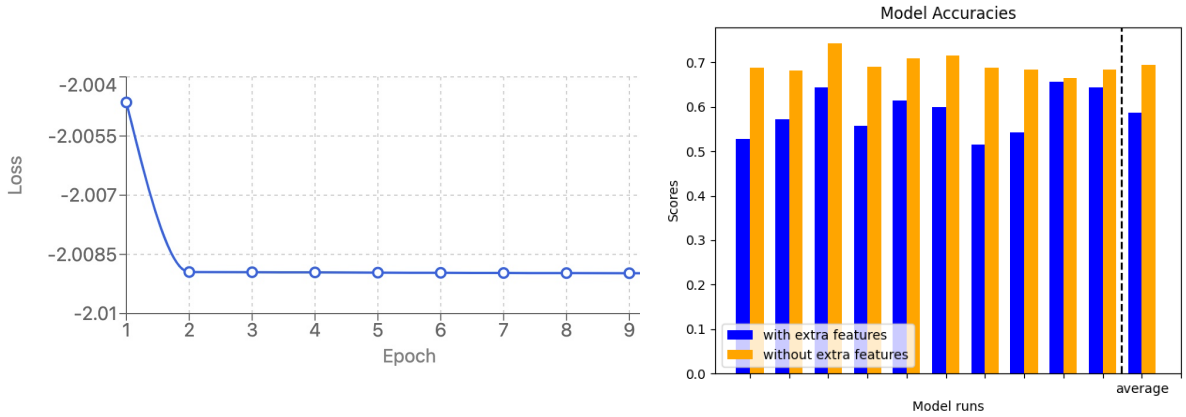

Figure 2: Transformer reconstruction loss over time; model accuracies, with and without extra features.

5 Discussion and Conclusion

To conclude, results concerning both cross-sectional and sequential data have shown great promise with respect to accuracy optimization. In particular, we show that our method of including additional features from the sequential data can improve the overall accuracy of the model by 10 percent. And while this method may seem trivial and obvious to use, its implications were not fully understood by our team until we performed quantitative comparisons.

In terms of the Kaggle competition, we are confident that our model will perform well, given the results of our cross-validation and the improvements made by learning on sequential data. However, we also understand that, due to time constraints and outside factors beyond our control, our model may not be the best and could use further improvements for medal qualification.

Lastly, while we planned to produce model accuracies on a feature-optimizing basis, we were unable to do so due to time constraints. We believe that this would have been a valuable addition to our project, and we hope to include this in future work. Architecture diagrams and relevant code can be provided upon request.

References

[1] Elias Aboujaoude. Problematic internet use: an overview. *World Psychiatry*, 9(2):85–90, June 2010.

[2] Hilarie Cash, Cosette D Rae, Ann H Steel, and Alexander Winkler. Internet addiction: A brief summary of research and practice. *Curr. Psychiatry Rev.*, 8(4):292–298, November 2012.

[3] Antonina Dolgorukova. Cmi-piu: Features eda, Nov 2024.

[4] Mauro Pettorruso, Stephanie Valle, Elizabeth Cavic, Giovanni Martinotti, Massimo di Giannantonio, and Jon E Grant. Problematic internet use (PIU), personality profiles and emotion dysregulation in a cohort of young adults: trajectories from risky behaviors to addiction. *Psychiatry Res.*, 289(113036):113036, July 2020.

[5] Anita Restrepo, Tohar Scheininger, Jon Clucas, Lindsay Alexander, Giovanni A Salum, Kathy Georgiades, Diana Paksarian, Kathleen R Merikangas, and Michael P Milham. Problematic internet use in children and adolescents: associations with psychiatric disorders and impairment. *BMC Psychiatry*, 20(1):252, May 2020.

[6] Adam Santorelli, Arianna Zuanazzi, Michael Leyden, Logan Lawler, Maggie Devkin, Yuki Kotani, and Gregory Kiar. Child mind institute — problematic internet use, 2024.

Appendix

A Description of Fields in Time-Series Dataset

Field	Description
step	Step count
X	X-axis acceleration of the heartrate monitor
Y	Y-axis acceleration of the heartrate monitor
Z	Z-axis acceleration of the heartrate monitor
enmo	Euclidean Norm Minus One (ENMO)
anglez	Angle in the Z-axis
non-wear_flag	Non-wear flag
light	Light exposure
battery_voltage	Battery voltage of the monitor
time_of_day	Time of day
weekday	Day of the week
quarter	Quarter of the year
relative_date_PCIAT	Current PCIAT minus previous day PCIAT

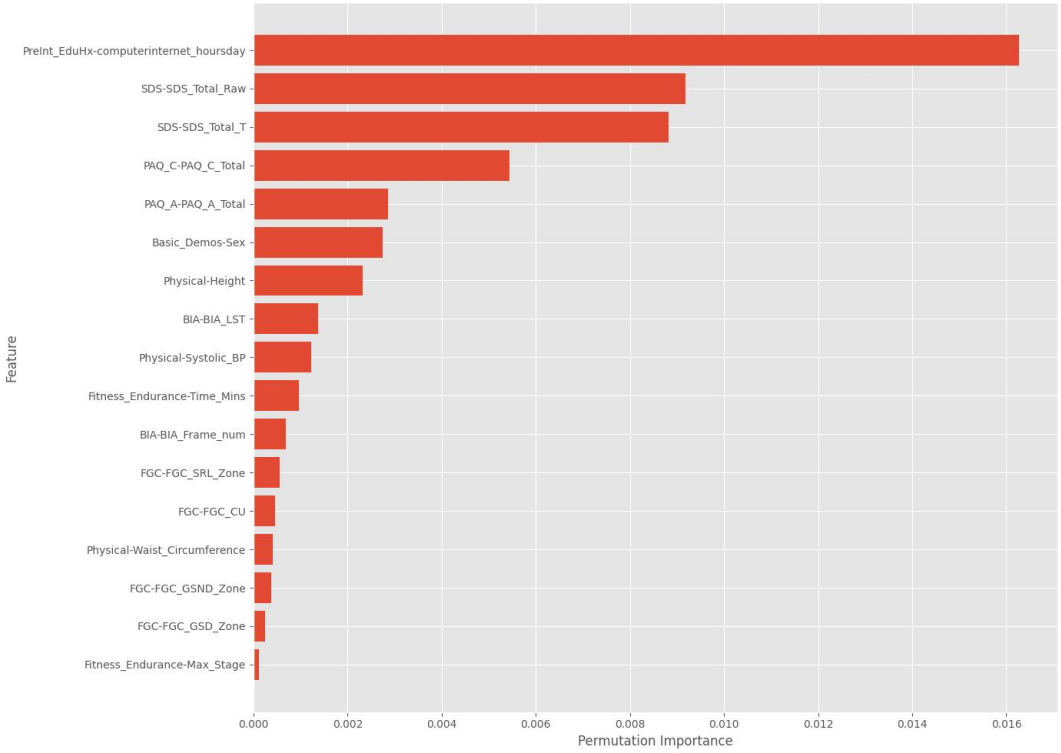

Figure 3: Features plotted in order of importance according to the XGBoost classifier, ascending.

B Accuracy of Feed Forward Neural Network on Cross-Sectional Data

Figure 4: Model accuracy of Feed-Forward Neural Network

C Hypothesis Testing

In order to compare two models A and B the null hypothesis is that the distribution of $acc(A)_i - acc(B)_i$ has zero mean and the alternative hypothesis is that the model with the higher mean performance accuarcy is significantly better than the other.
T statistic is given by:

$$t = \frac{\overline{acc}(A) - \overline{acc}(B)}{\sqrt{var(A-B)/k}}$$

where,

$$var(A-B) = \frac{1}{k}\sum_{i=1}^{k}[acc(A)_i - acc(B)_i - (\overline{acc}(A) - \overline{acc}(B))]^2$$

Model 1	**Model 2**	**t-statistic**	**p-value**	**Accept/Reject Null**
RF	XGB	-0.492	0.633	Accept
RF	SVC	6.149	0.0	Reject
RF	FFN	-0.039	0.970	Accept
XGB	FFN	0.304	0.767	Accept
XGB	SVC	4.104	0.002	Reject
FFN	SVC	4.588	0.001	Reject

Table 3: Comparison of Models using t-statistic and p-value

Therefore, we can conclude that RF, XGB and FFN are significantly better than SVC. However it looks like the distribution of accuracies of RF , XGB and FFN are comparable. In this case we will choose the model that requires the least amount of parameters which is the XGBoost classifier.