Child Mind Institute — Problematic Internet Use

UNIVERSIDAD DISTRITAL
FRANCISCO JOSÉ DE CALDAS

UNIVERSIDAD DISTRITAL FRANCISCO JOSÉ DE CALDAS

Integrants:

Jan Henrik Sanchez Jerez – 20231020130

Sebastian Villarreal Castro – 20221020059

Juan David Quiroga - 20222020206

System Analysis

Presented to:

Carlos Andres Sierra Virguez

2025

Summary

Our system analysis revolves around the study of the System Impact Index (SII), which aims to quantify the deterioration of children's mental health associated with problematic internet use. To achieve this, we work with a combination of demographic, physiological, and behavioral data obtained from both the **Parent-Child Internet Addiction Test (PCIAT)** clinical questionnaire and physical monitoring devices (actigraphy). This fusion of data sources—each with its own limitations and characteristics—gives rise to a system that is inherently complex and loaded with uncertainty.

In predictive terms, the core goal of the system is to estimate the severity of problematic internet use in children and adolescents, represented by an integer value ranging from 0 to 4. This value serves as the output label in a supervised multiclass classification task. However, while these numbers reflect increasing severity levels (e.g., 0 indicates low severity and 4 high), traditional classification models do not interpret this ordinal relationship. To the model, the classes behave as independent categories (akin to colors: red, green, blue), with no understanding that an error between 3 and 4 is more acceptable than one between 0 and 4. This is a critical point that affects both model design and the choice of evaluation function.

Throughout our analysis, we identified a considerable proportion of missing values, particularly in physiological data (such as heart rate and sleep parameters) collected via wearable sensors. This lack of information represents a structural constraint of the system, as it impacts both imputation processes and the feasibility of supervised training, potentially introducing severe bias if ignored. To address this issue, we propose robust preprocessing strategies such as median imputation (less sensitive to outliers) and temporal interpolation, to restore incomplete trajectories without drastically altering the dynamics of the data.

Additionally, we face a dual challenge regarding data quality: subjectivity and temporal inconsistency. On one hand, the **PCIAT questionnaire**, although clinically validated, is susceptible to response bias from parents or caregivers. On the other hand, actigraphy data exhibit imbalances in measurement duration, gaps due to device non-use, and temporal synchronization errors, all of which introduce noise that is difficult to eliminate. This landscape suggests that the data are governed by non-deterministic and potentially chaotic processes, where small initial differences may be amplified over time. This observation aligns with principles from chaos theory, which describes systems where long-term behavior is highly sensitive to initial conditions, making deterministic predictions unreliable. In this context, the system under study resembles a chaotic dynamic system, in which even slight perturbations can lead to disproportionate effects on the trajectories of the physiological and behavioral variables being analyzed.

From a predictive modeling perspective, we identify two types of feedback that govern the system:

- **Internal feedback** during supervised training, based on metrics such as cross-validation.

- **“External feedback”**, derived from the use of the **Quadratic Weighted Kappa (QWK)** coefficient as the final evaluation metric. The QWK is particularly relevant here, as it penalizes large errors in ordered categories (e.g., predicting a 0 instead of a 4) more heavily than minor ones (e.g., predicting a 3 instead of a 4). Given this, we believe that QWK is not only useful as a final metric but could also be employed as a loss function during model training, making it more sensitive to the ordinal nature of the target variable.

Taken together, the results of the analysis reveal that we are working in a highly heterogeneous, incomplete, and noisy data environment, where any robust predictive design must consider the following key constraints:

- Frequent missing data in physiological signals.
- Subjectivity in the PCIAT questionnaire responses.
- Temporal noise, asynchrony, and variable duration in the recordings.
- The ordinal nature of the target variable, which must not be ignored by the model.
- The presence of chaotic or nonlinear dynamics, especially in temporal signals, which calls for tools capable of modeling uncertainty and dynamic complexity.

Given this context, we conclude that any predictive architecture we propose must incorporate robust preprocessing techniques, probabilistic modeling elements, bias-mitigation strategies, and evaluation metrics tailored to the ordinal nature of the problem.

2. Design Requirements Derived from the Analysis

Based on the analysis of the System Impact Index (SII), we identify the following design requirements that should guide the development of the predictive system:

Prediction quality as a top priority: Since the system’s primary goal is to estimate the severity of problematic internet use, model accuracy is prioritized over execution speed. Model performance should be evaluated primarily using the Quadratic Weighted Kappa (QWK) metric, which appropriately reflects the ordinal nature of the classification problem.

Robust handling of missing data: Given the high proportion of missing values—particularly in physiological signals—the system must include resilient preprocessing techniques, such as median imputation and temporal interpolation, to reconstruct incomplete time series without significantly altering their internal dynamics.

Fault and noise tolerance: Due to inherent noise in the data (temporal inconsistencies, desynchronization, and potential chaotic behavior), the model must be robust to small input perturbations and measurement errors.

Ordinal-aware modeling: The model must explicitly account for the ordinal structure of the target variable. This includes not only evaluating with ordinal metrics like QWK but also adopting loss functions that penalize large misclassifications more heavily than minor ones.

Probabilistic adaptability: Given the presence of uncertainty and non-determinism, the system should incorporate probabilistic or Bayesian modeling elements to capture subject-level variability and temporal instability.

Reasonable computational efficiency: Although real-time execution is not required, the system should maintain acceptable inference times and avoid being excessively computationally expensive, ensuring practical usability in clinical or research settings.

- User-Centric Needs

While the system is primarily data-driven and analytical, certain user-centered aspects should be considered to ensure broader accessibility and trust:

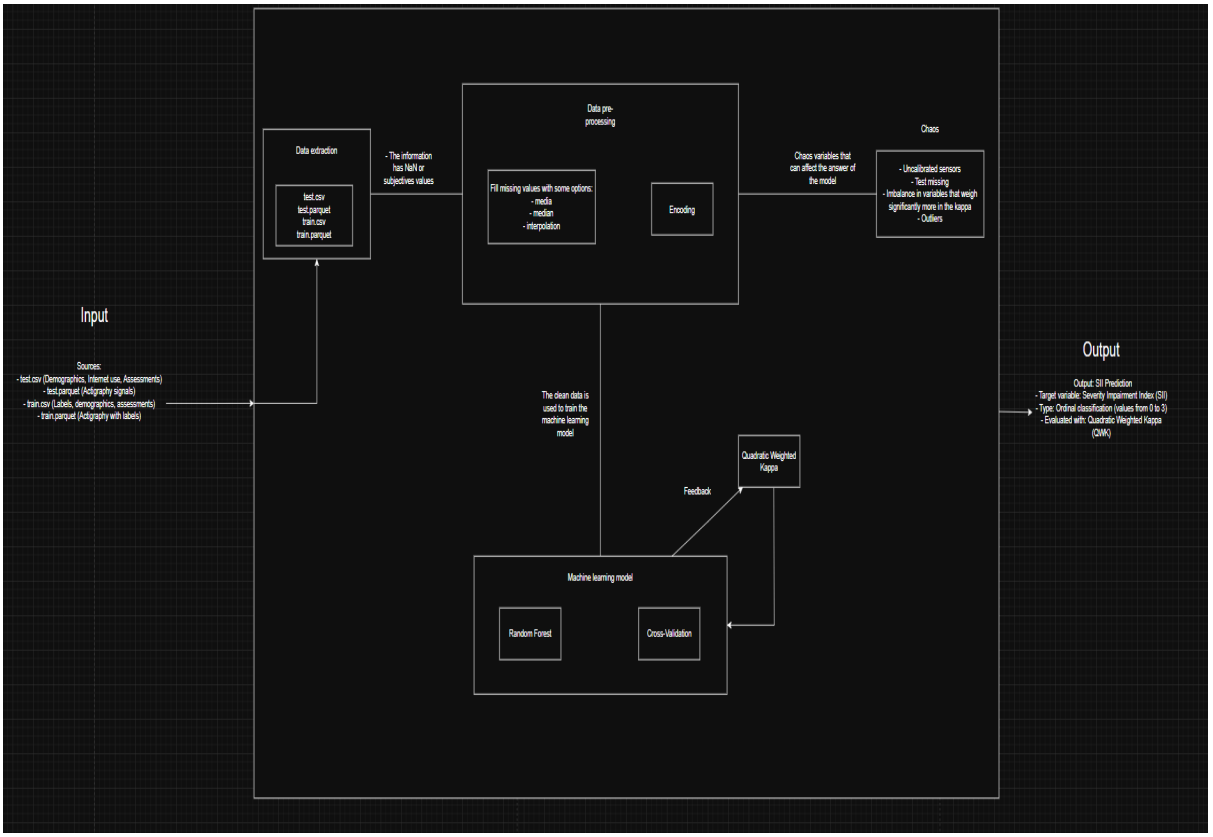
Interpretability of outputs: The system must present predictions in a clear and understandable format for non-technical users, such as psychologists or pediatricians. This includes the use of simple and intuitive interfaces, visual summaries, and accessible explanations of results. In this context, interpretability refers to the ease of understanding the predictions and their implications, not necessarily the internal logic of the model.

Model transparency (if required): In more technical or clinical scenarios, the system may also benefit from offering insights into how predictions are made. This could be achieved through inherently interpretable models (e.g., decision trees) or post-hoc explanation methods like SHAP or LIME, especially if trust and accountability are important in decision-making contexts.

Basic usability and accessibility: While ease of use is not a core requirement, offering a user-friendly interface improves adoption and reduces friction. The system should provide structured outputs, handle invalid or incomplete inputs gracefully, and offer feedback that helps users understand any issues with their data.

System reliability: Even though no highly sensitive or private data is processed, the system should be resilient to faults, capable of operating smoothly in the presence of data irregularities, missing values, or brief interruptions in the input signals.

3. High-Level Architecture:



1. Input: Data Sources

The system begins by collecting data from four primary sources:

- **test.csv**: Contains demographic information, internet usage data, and questionnaire assessments (no target labels).
- **test.parquet**: Includes actigraphy signals recorded by physical monitoring devices.
- **train.csv**: Same structure as test.csv, but includes the target variable (SII).
- **train.parquet**: Actigraphy signals aligned with corresponding target labels.

These files combine structured tabular data and temporal sensor data. This input is fed into the **Data Extraction** module.

2. Data Extraction

This stage loads and merges the various input sources. It also identifies common issues within the dataset, such as:

- Missing values (NaN)
- Subjective or non-numeric variables

Recognizing these issues early is essential for effective pre-processing and modeling.

3. Data Pre-processing

The data pre-processing step ensures the dataset is clean, consistent, and usable for machine learning models.

Key actions include:

- **Missing Value Imputation**: Replacing absent data using techniques such as:
    - Mean
    - Median
    - Interpolation (especially useful for time-series data like actigraphy)
- **Encoding**: Transforming categorical or qualitative features into numerical representations suitable for model input.

Once cleaned, this data is used to train the machine learning model. However, at this point, the system must also consider the presence of unpredictable or problematic variables—referred to here as **chaos**.

4. Chaos

"Chaos" represents external or internal data inconsistencies that can interfere with model training or predictions. These include:

- **Uncalibrated sensors**: Hardware issues may produce inaccurate actigraphy data.
- **Missing tests**: Some participants may have incomplete assessments or questionnaires.
- **Imbalanced variables**: Certain features might disproportionately influence predictions or the evaluation metric.

- **Outliers**: Extreme or abnormal values that can distort the learning process

These issues do not follow a predictable pattern and are difficult to eliminate completely. Their presence reinforces the need for robust model training and validation procedures.

5. Machine Learning Model

The model chosen is:

- **Random Forest**, due to its:
    - Tolerance for noisy data
    - Ability to handle complex and non-linear relationships
    - Interpretability and performance with heterogeneous data
- **Cross-validation** is implemented to test the model across different data subsets and ensure generalization. This is particularly important given the possible imbalance and presence of outliers in the data.

6. Evaluation: Quadratic Weighted Kappa (QWK)

The model's performance is evaluated using **Quadratic Weighted Kappa**, a metric designed for **ordinal classification tasks**. Unlike typical accuracy or F1-score metrics, QWK considers how far off the prediction is from the actual label, penalizing larger deviations more heavily.

This evaluation process provides **feedback** to fine-tune the model or improve pre-processing steps.

7. Output: SII Prediction

The final goal of the system is to predict the **Severity Impairment Index (SII)** for each subject. The output characteristics are as follows:

- **Target variable**: Severity Impairment Index (SII)
- **Type**: Ordinal classification, with values ranging from 0 to 3
- **Evaluation metric**: Quadratic Weighted Kappa (QWK), selected for its sensitivity to misordered classifications

4. Addressing Sensitivity and Chaos:

The design recognizes that the data set is subject to significant levels of noise, inconsistency, and subjectivity, especially in physiological and behavioral measurements. These characteristics introduce chaotic behavior into the system, where small variations in the input data such as missing sleep data or gaps in actigraphy can lead to disproportionate changes in model predictions. To manage these problems, the following strategies have been implemented:

Working with High Sensitivity Variables

- Ordinal-aware modeling: Since the target variable SII is ordinal, we adopt QWK evaluation and potentially loss functions that respect the distance between class labels, reducing the impact of high misclassifications.

- Cross validation: Stratified k-fold cross validation is used to ensure performance consistency across variable distributions, accounting for class imbalance and minimizing model variance.

Sources of Chaos and Their Impact

- Much of the information has NaN or subjective values: Missing values reduce data completeness and compromise statistical power. Subjective data, like questionnaire responses, introduces bias and noise due to individual perception differences.

- Uncalibrated sensors: Measurements from physical devices may be inconsistent or unreliable, especially across participants. This affects the stability and comparability of features like ENMO or light levels.

- Test missing: Incomplete clinical tests missing PCIAT or treadmill results create gaps that cannot always be recovered through imputation, reducing the ability to model certain behaviors accurately.

- Imbalance in variables that weigh significantly more in the kappa: When some variables extreme IBS cases, the model may underfit these classes. This leads to disproportionately high score penalties when misclassified, and an overall degradation of the model's fairness and robustness.

Some ways to mitigate the chaos

- Feedback Loop Monitoring: The system evaluates model outputs using the Quadratic Weighted Kappa metric and feeds the score back into the training pipeline. This iterative refinement allows dynamic adjustment based on model behavior across ordered categories.

- Noise tolerant Algorithms: Algorithms like Random Forest are chosen for their robustness to noisy or partially missing data, and their ability to model non-linear

interactions without overfitting.

- Subjectivity damping: PCIAT scores derived from perceptions are supplemented with objective features derived from actigraphy (e.g., mean ENMO), reducing the exclusive reliance on subjective sources.

5. Technical Stack and Implementation Sketch:

Proposed Tools and Frameworks

- Programming Language:
  - Python 3.11
- Data Handling:
  - pandas
  - polars
  - pyarrow
- Numerical and Feature Engineering Libraries:
  - numpy
  - scikit-learn
  - tsfresh
  - pyts

- Modeling Framework:
  - pytorch

- Execution Environment:
  - Google Colab

Justification of Selected Technologies

- Python 3.11:

  Chosen for its modern syntax improvements and wide compatibility with machine learning libraries. Python remains the standard language for data science and ML development.

- pandas / polars / pyarrow:

  These libraries offer fast, flexible, and memory efficient tools for reading and transforming tabular and time series data. pandas is the go for general manipulation, polars enhances performance on large datasets, and pyarrow enables efficient reading

of parquet files from actigraphy data.

- numpy / scikit-learn:

  Fundamental for numerical operations and traditional machine learning preprocessing. scikit learn also provides tools for cross validation, metrics like QWK, and pipeline design.

- tsfresh / pyts:

  Designed to extract time series features for example from actigraphy data. These libraries allow for statistical summarization and transformation of sequential sensor inputs into fixed size feature vectors usable by tabular models.

- pytorch:

  Used for implementing deep learning models, including 1D CNNs or RNNs to handle sequential data directly. PyTorch's flexibility and support for custom loss functions make it ideal for integrating ordinal-aware losses such as QWK-based optimization.

- Google Colab:

  Provides a cloud based Jupyter environment with free access to GPUs and sufficient memory for both exploratory data analysis and model training. It supports collaborative work and eliminates local hardware limitations.

References

https://www.kaggle.com/code/reighns/understanding-the-quadratic-weighted-kappa

https://www.healthyplace.com/psychological-tests/parent-child-internet-addiction-test