Workshop 2

1st Bettsy Liliana Garces Buritica
code: 20231020222
blgarcesb@udistrital.edu.co

2nd Marta Isabel Sanchez Caita
code: 20222020118
maisanchezc@udistrital.edu.co

3rd Luis Fernando Rojas Rada
code: 20222020242
lfrojasr@udistrital.edu.co

4th Mauricio Daniel Baes Sánchez
code: 20222020058
mdbaess@udistrital.edu.co

I. Summary of Analytical Findings

The analysis of the Child Mind Institute – Problematic Internet Use challenge describes a system that operates at the intersection of human behavior, digital interaction, and physiological activity. It can be understood as a socio-technical system that connects three main components: (1) biological signals, such as movement and sleep cycles captured through actigraphy; (2) behavioral and cognitive factors, reflected in questionnaires and psychometric assessments; and (3) environmental and contextual variables, including screen exposure patterns and lifestyle indicators. Together, these components generate a complex network of relationships that evolve over time and are influenced by both internal (biological and psychological) and external (technological and social) dynamics.

This system is characterized by nonlinearity and high sensitivity to change. Small variations in physical activity, sleep quality, or emotional state can lead to disproportionate shifts in behavior and internet use patterns. For example, a few nights of poor sleep may alter daily energy levels, which in turn affects activity routines and digital engagement. These interactions reflect the type of feedback loops commonly found in adaptive human systems: reinforcing loops that amplify problematic use (e.g., fatigue leading to more screen time) and balancing loops that help regulate it (e.g., social interaction or physical activity reducing screen dependence). Such loops make the system inherently dynamic and difficult to predict over long periods.

A second defining feature is data heterogeneity and incompleteness. The dataset includes diverse types of information — continuous actigraphy signals, categorical survey responses, and ordinal mental health indicators — that operate on different temporal and measurement scales. Approximately one-third of the main outcome variable is missing, and the distribution of categories is highly uneven, creating additional systemic constraints. These conditions introduce uncertainty and make it challenging to isolate direct causal relationships.

From a systems perspective, this environment behaves as a complex adaptive system: it evolves through interactions among components rather than linear cause–effect chains. The data capture only partial snapshots of ongoing processes influenced by attention, emotion regulation, and external stimuli such as device use or social context. Because of this, any predictive approach must consider not just statistical patterns but also the dynamic dependencies that underlie them.

II. Definition of System Requirements

A. Functional Requirements

The system must take into account two types of data: CSV files containing clinical and tabulated information on participants, such as their age, gender, anthropometric measurements, fitness tests, body composition, internet usage patterns, and sleep scales, and PARQUET files of data from motion sensors captured with wrist accelerometers, containing up to 30 days of daytime and nighttime activity (3-dimensional acceleration, movement intensity, angle, ambient light, voltage, periods of non-use, etc.).

Due to the large amount of missing data and particularly the fact that approximately 30% of severity labels are absent, the system will need to implement robust and intelligent imputation techniques. This is essential to ensure that learning, modeling, and predictions are not compromised as a result.

From this data, the system must extract meaningful characteristics: from the usual clinical indicators (BMI, blood pressure, physical condition) to metrics that can assume physical activity and sleep, such as average movement intensity, duration of active and inactive periods, circadian rhythms, etc. It must also pay attention to the temporal relationship between movement data and the administration of the questionnaire that assesses problematic internet use in order to better understand the sequence of events.

The central element is an ordinal regression model, which classifies each participant according to one of four possible levels of severity: none, mild, moderate, or severe. To evaluate the quality of the predictions, the Quadratic Weighted Kappa metric is applied, which is a metric that penalizes errors more the further the predicted level is from the actual level. The system must support pre-trained models, ideally based on algorithms such as XGBoost, LightGBM, or CatBoost, but also neural networks, ensuring good performance based on cross-validation.

The interface must be a simple, user-friendly web page so that users can easily upload files, view the predicted result with the model's confidence level, and see an interpretive explanation highlighting the importance of the variables that have the greatest impact. Finally, the system must be able to generate downloadable reports in CSV format containing the participants' identifiers and their predictions, ready for competition or subsequent analysis.

B. Non-functional Requirements

The system must be efficient with large volumes of data, reasonably fast in obtaining results from individual entries (always under 30 seconds per prediction), and capable of handling multiple batch entries in a reasonable amount of time. The architecture must be modular to accommodate future improvements and must be able to run on multiple devices, either through an interface that responds to the appropriate format of mobile device screens (responsive) or through web access.

Although authentication will not be implemented for version 1, minimum security for transfer and temporary storage must be ensured by implementing encryption and file deletion policies after a reasonable period of time, thus protecting privacy.

Priority is given to a simple user experience (without technicalities), with explicit and useful messages in case of errors, and with all the documentation explaining everything from installation to use and troubleshooting.

Finally, the system must be reliable, with strong input validation, adequate error or exception handling, activity logs to facilitate system maintenance, and the ability to recover from failures or interruptions (at a minimum).

C. Technical Requirements of the Model

The model must understand that the severity of problematic internet use is ordinal (a severe prediction error when there is none is more severe than confusing mild with moderate). Therefore, it must use ordinal regression with the ability to also provide uncertainty estimates.

The model ecosystem must efficiently integrate tabular data and actigraphy time series, either through feature engineering or hybrid models. The main metric will be Quadratic Weighted Kappa, as well as complementary analyses of precision, recall, or F1 for each class to better understand the model's behavior at each level.

The architecture must allow for easy model exchange, initially using pre-trained models for speed, but with the possibility of training your own models when more time or data is needed. The inference API must be standard and easy to share, to facilitate integrations and future updates.

III. High level architecture

1) **Data upload:** The user uploads a `CSV` file with the tabular data and a `Parquet` file with the accelerometer series.
2) **Preprocessing:** The backend calls the data processing module to process the files (cleaning, normalization, feature extraction).
3) **Prediction:** The model runs the inference and returns a probability per class (*None*, *Mild*, *Moderate*, *Severe*).
4) **Storage:** The results are saved in the database along with the patient ID, all stored in the cloud.
5) **Display:** The dashboard shows graphs with distributions, alerts, or clinical reports.

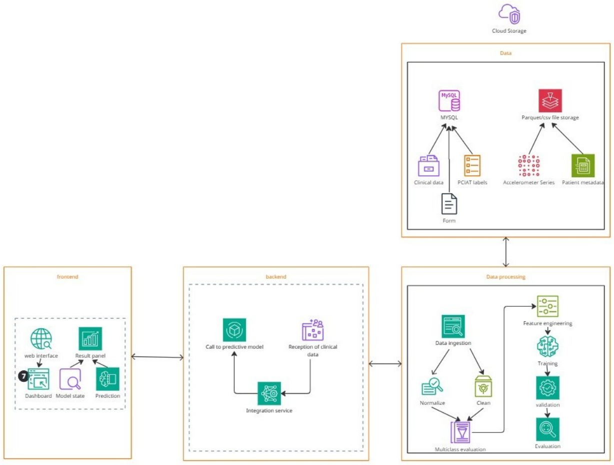

Fig. 1. Architectural diagram

Principles and their Application to the Design

A. Modularity

Dividing the system into independent, cohesive components reduces complexity and improves maintainability.

- **Decision:** Separation into layers (Frontend / Backend / Data Processing / Infrastructure).
- **Benefit:** Each module can be developed, tested, and deployed independently (e.g., update the predictive model without changing the web interface).

B. Availability and Resilience

Minimizing single points of failure and ensuring continuity in case of system errors.

- **Decision:** Distributed responsibilities and deployment in managed infrastructure (buckets, databases, gateways).
- **Benefit:** Faster recovery and fault tolerance (e.g., restarting the inference service without data loss).

C. Validation and Verification

Ensuring correctness of inputs, processes, and outputs.

- **Decision:** Input schema validation, automated integration tests (backend–model), and evaluation metrics for predictions.
- **Benefit:** Prevents ingestion of corrupted data and guarantees reliable results for clinical interpretation.

Conclusion

Systems Engineering principles guided the design decisions of this architecture. Modularity and separation of concerns justified the multi-layer structure; scalability and resilience motivated the adoption of cloud-managed services; and principles of security, observability, and data governance shaped the supporting infrastructure. Applying these principles ensures the system's robustness, maintainability, and reliability in a clinical context.

IV. Sensitivity and Chaos

The predictive system designed to identify problematic internet use (PIU) operates within a highly dynamic and nonlinear environment. Small variations in input data—such as movement intensity, sleep patterns, or self-reported emotional states—can produce disproportionate effects on the model's outputs. This sensitivity reflects both the complexity of human behavior and the instability inherent to multimodal datasets collected in real-world contexts. Therefore, the system's design incorporates strategies to reduce the influence of chaos and enhance its adaptability.

First, data preprocessing and normalization routines are applied to minimize the amplification of random fluctuations. Actigraphy time-series data undergo smoothing and noise-filtering processes, while categorical and psychometric variables are standardized to maintain comparable scales. This reduces sensitivity to extreme values and promotes model stability during training. Additionally, intelligent imputation techniques are used to manage missing or inconsistent data, preserving temporal and behavioral coherence.

Second, the system integrates adaptive recalibration and feedback mechanisms. Since behavioral and emotional indicators interact through feedback loops (for instance, lower physical activity leading to increased screen time, which then affects sleep quality), the model architecture seeks to capture such dependencies. By incorporating historical and lagged variables, the system can distinguish between short-term anomalies and consistent behavioral shifts. Periodic re-training and validation cycles allow the model to adjust to new data, ensuring robustness as population characteristics evolve.

Third, the design includes monitoring and error-handling mechanisms to address unforeseen conditions during operation. These routines aim to detect anomalous data deviations or unexpected model behavior—such as sudden changes in variance or inconsistent relationships among variables. Their purpose is to strengthen system stability and support corrective adjustments when necessary, maintaining resilience without overreacting to noise or uncertainty.

Considering that the system models inherently chaotic human behaviors, interpretability and transparency are prioritized over mere statistical accuracy. Through interpretability techniques such as feature attribution or dependency mapping, the system clarifies how variables like sleep duration, screen exposure, or emotional variability contribute to the risk of problematic use. This interpretability serves as a stabilizing factor by enabling human oversight and contextualizing unexpected patterns, ensuring that the system remains robust, adaptable, and consistent with the unpredictable nature of behavioral and physiological data associated with problematic internet use.

V. Technical stack and implementation diagram

A. Technological stack

During the planning stage of development of the website it's necessary that the team agrees in which technologies to use, for this specific problem the necessary goals that the technological tools need to accomplish are as follows:

Programing language(s):
The chosen language or language has to have the necessary tools for front end and back end web development and the means to adapt a website for most common hardware (pcs, phones, tablets) some viable options are:

JavaScript, Pyton, C++, Chrome WebTools

The website is planned to be mostly written in Python for the back end and JavaScript for the front end, but other languages may be used depending on the requirements and the team's knowledge.

IDE:
Necessary for programming, mostly comes down to personal preferences but the easiest to access that can handle most languages is:

Visual Studio Code

SQL database management tool:
With SQL being the Database management language that most of the development team has experience with and the necessity to store form answers, prediction results, and user feedback, a SQL server manger app its required, some options are:

Beaver DB MySQL
With MySQL being chosen for the task.

Class Categorizing AI model:
The pre-trained model that will be used to process the information given by the client, predicts the relation between the subject's behavior and internet usage and categorize the result in one of the four options.
Open AI's GPT Will be the model used to do this, chosen due to its accessibility and previous training.

B. Implementation sketch

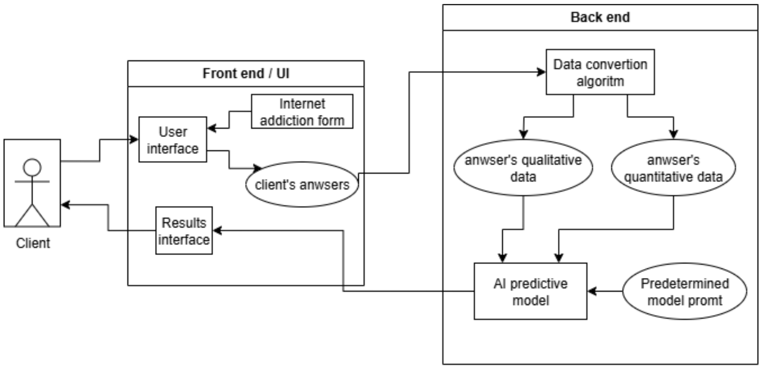

Fig. 2. Implementation sketch

References

[1] Child Mind Institute, “Child Mind Institute — Problematic Internet Use,” *Kaggle Competition Platform*, 2024. [Online]. Available: https://www.kaggle.com/competitions/child-mind-institute-problematic-internet-use