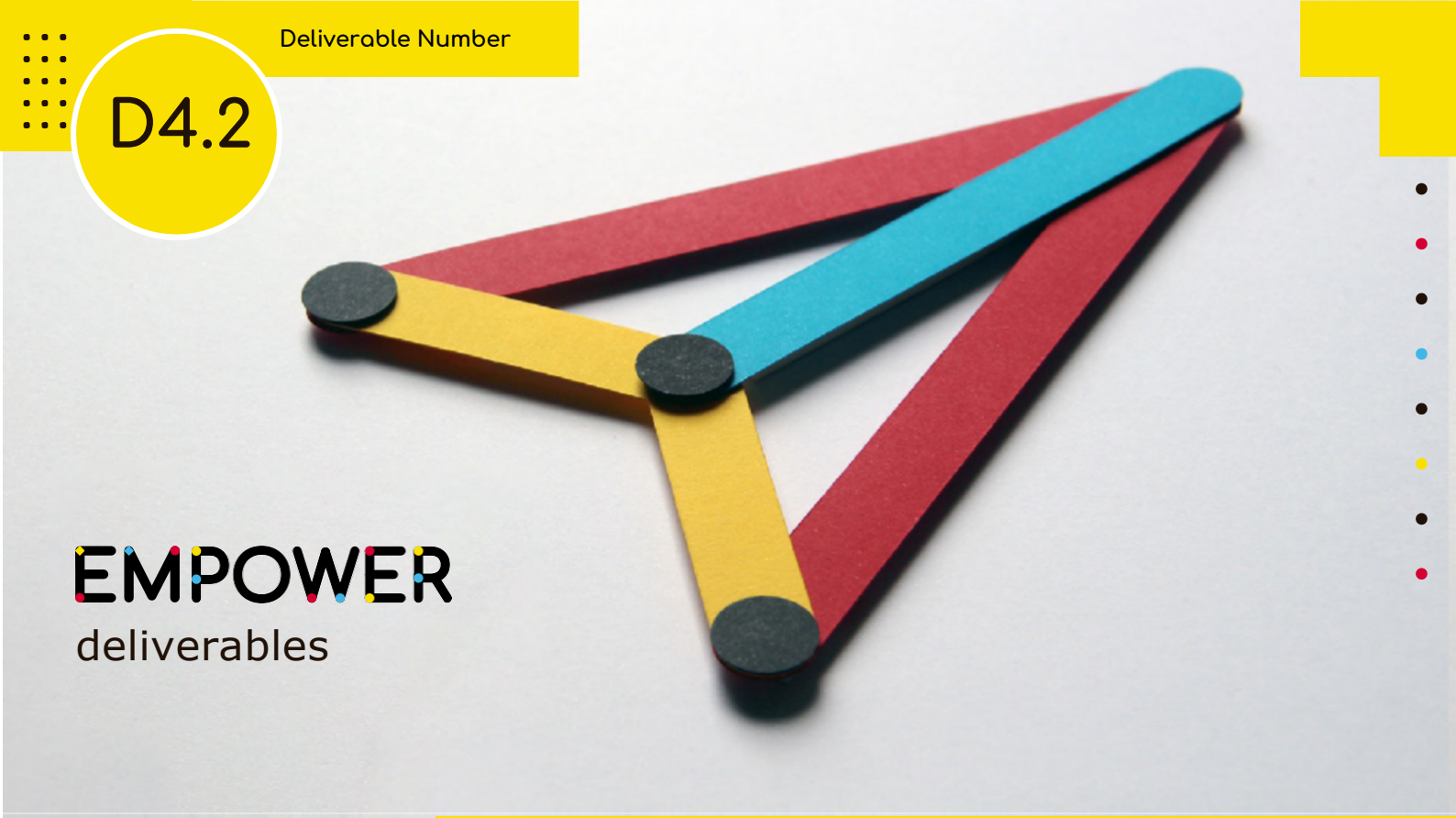
Deliverable Number

D4.2



EMPOWER

deliverables

Deliverable name	**INITIAL ALGORITHMS**
Type	R - Document Report
Dissemination level	PU - Public
Date	Month 12

A report on the initial algorithmic development demonstrating the ability to identify relevant aspects as defined in D4.1 from sensors used on the platform in WP3

Description

WP.4

Work Package. 3

Lead Beneficiary – RU

Initial Algorithms

Executive Summary

This document describes recent work developed by EMPOWER WP4 regarding the initial algorithmic development, demonstrating the ability to identify relevant aspects as defined in D4.1 from sensors used on the platform in WP3.

Date	Version	Description	Authors
14.09.2023	0.1	First Draft: initial structure of the document	Joana Campos, Marcos Bueno, Joana Brito
29.09.2023	0.2	Description of the conceptual AI approach and empirical results of initial machine learning models	Joana Campos, Marcos Bueno, Ana Paiva, Joana Brito

Table of Content

#1. INTRODUCTION *4*
#2. A hybrid Approach to Algorithm Development **5**
#3. Example: Working Memory Game **6**
#4. Generalizations **9**
#5. Preliminary empirical results **9**

#1. INTRODUCTION

This deliverable describes the approach used for modelling gaming elements, cognitive aspects, and predictions. This is a hybrid AI approach that combines data with domain knowledge. We first describe the hybrid approach and illustrate it concretely by means of an example, namely the Working Memory game.

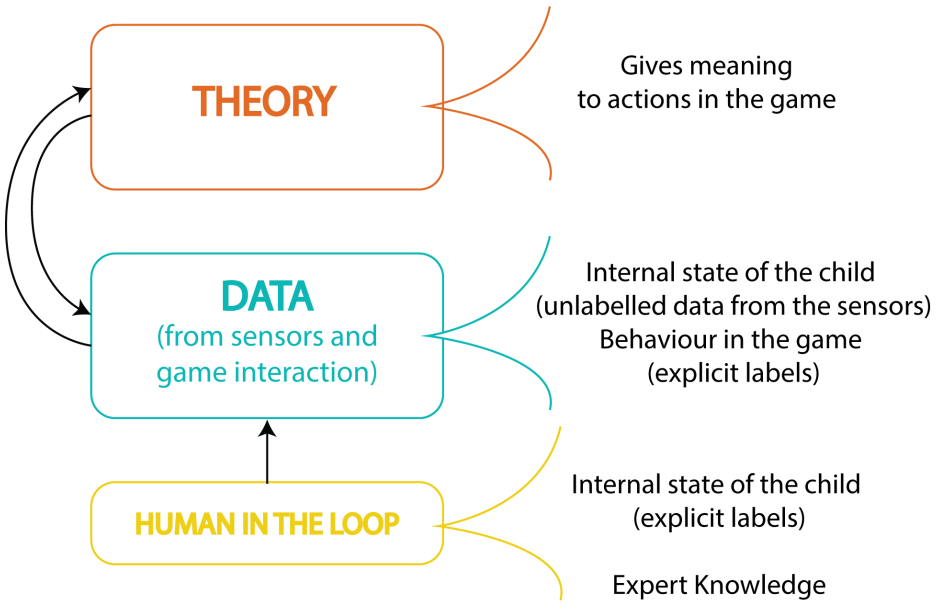
In the final part, we discuss initial predictive models build from a pilot-study data, which shows initial empirical results. These results are also useful for the data collections in the upcoming pilot studies to maximize the predictive power of the proposed AI approach.

#2. A Hybrid Approach to Algorithm Development

Given the difficulties of obtaining reliable data from the sensors at this stage, we outline a natural direction to the task. In this project where the algorithms will rely on different types of information, but still operate in the absence of one data stream (e.g., wearables, eye tracking, etc.).

For that, we will consider a ***hybrid approach*** (Figure 1) for creating the game's assessment tools and action elements. A hybrid approach refers to a model that combines machine learning for certain aspects of the system and hand coded actions for other, more accessible aspects that the system will execute. A hybrid approach considers that we will not only rely on data and algorithms created with the data collected in the initial data collection studies but also extract elements and procedures that will be modelled as rules for assessment and action in the system(see Table X).

The approach will generally have three components: Theoretical concepts, Data Collected,



Experts (Human-in-the-loop).

Theoretical Concepts are used to map cognitive functions to the game mechanics of the developed games. Theory informs events of interest in the game and gives meaning to the child's behaviour. Overall, it provides the base rules for performance within each game. Tables 1 and 2 illustrate the game mechanics and possible game interactions for the Working memory game.

Data that comes from the sensors capture the internal state of the child. At the beginning we don't know what signals variations mean. The data training process uses the expert annotations game data (already semantically distilled according to the theory) as labels to make predictions. Those include game adaptations and feedback to the stakeholders. As described in Tables 1 and 2, a lot of interesting data for assessing student performance comes from the way the child interacts with it.

Following a hybrid approach, such data coupled with expert and theoretical knowledge could already produce a rule-based AI module encompassing game (explicit) behaviour. Within this framework, this layer provides markers and feedback to the sensor layers (in a continuous feedback loop), thereby improving classification and prediction.

Expert knowledge is intended to increase the robustness of the algorithmic component, the experts also provide a set of rules encompassing behaviour and predictive knowledge based on theory (in an initial phase) and observed in the initial studies (when that data is available).

Furthermore, experts should be able to decide whether the algorithms' decisions must be revised and help retrain the AI models. The role of the expert could be taken further, by asking teachers to evaluate in context several variables of interest (using the *Teacher App*). Such assessments could be used to add labels to the signal data.

#3. Example: Working Memory Game

Using the Working Memory Game, this section intends to illustrate how AI algorithms leverage the hybrid approach to provide adequate predictions and classifications.

We identify a set of *Game Mechanics (GM)* and *Player Actions (PA)*. *Game Mechanics* are the rules that govern the player's actions, and in the context of this project, we will refer to them as the mechanisms that are part of the game's playable elements with a direct link to the executive function of interest. *Player Actions* refer to interactable behaviours within the game.

The working memory game is designed to assess a child's memory span. This game is divided into two phases. First, children must **sort (GM1)** the good (yellow peppers) from the bad peppers (yellow peppers with a worm) so none will be wasted, as they turn yellow on the screen. While doing this, they must remember their **order of appearance** to **pick up the ripe peppers (GM2)** in a subsequent phase. Each level of the game has five trials. (Refer to Deliverable 3.1 for a more detailed game description).

The sorting task intends to add a distractor to the task of remembering the order of pepper

appearance. Although it does not have a direct role in recall performance, it increases the load of information and makes recall more challenging.

For this game, there one variable of interest: Length Span (volume of working memory) or Cognitive Load, whose measure is supported by several aspects designed in the game, as described in the Tables below.

Table 1 - Description of the cognitive meaning of each Game Mechanic in the Working Memory Game, how it links to a measure of success and how the signals could augment the performance of the AI algorithms.

Game Mechanics	Cognitive Meaning	Link to task success	Useful signals
Sorting good from bad peppers	Good or bad recall (increases load of information)		Eye Tracking (check whether error it is boredom)
Peppers order of appearance	Ability to recall information		Eye Tracking helps to follow the recall process.
GM3	Number of trials within levels	To make sure it was not by chance/random (it is accordance with the standardised task)	3 levels of success (still not sure the cut-off point) in bigger samples look at the mean, but now we considering the individual.	Heart rate and Eye tracking can give an estimate of attention and how easy the task was.
GM4	Number of peppers to recall (level)	According to the standarized task (length span)	Longest sequence they recall	Heart rate and Eye tracking can give an estimate of attention and how easy the task was.
GM5	Interference (related to GM1)	Interference to the recall activity. If they fail to recall all but have done better before, the interference is playing a role.		
GM6	Distance between	Closer are	Positions	

	consecutive Peppers	easier to remember; intersection of the paths	correlated with task performance	

Table 2 - Description of the cognitive meaning of each Player Action in the Working Memory Game, how it links to a measure of success and how the signals could augment the performance of the AI algorithms.

ID	Player Actions	Cognitive Meaning	Useful signals
PA1	Pepper selected in order	Recall	
PA2	Correct Pepper, but not in the correct order	Different levels of recall (expert knowledge required to create explanations)	
PA3	Time to select correct pepper	Processing speed (may correlate with sustained attention)	
PA4	Time to select wrong pepper	Processing speed, trying to process the information, but there exists some interference.	Heart rate and sustained attention
PA6	Follow (with eyes) pepper appearance	Strategy and task performance	ET tracking

#4. Generalizations

We envision that the AI algorithms will support the flow of each game independently, as many of the predictions and classifications are game-dependent. However, at a high-level layer, the AI algorithms can provide generalisations by looking at the aggregated users and transversal elements in all games.

Aggregated Users data provides the AI algorithms with in-game patterns in strategies. It will allow us to answer questions such as *What is the best strategy? What is the best training script for children with a confident performance?*

At the system level, the system could predict whether the child will be challenged by a different level or whether the level is too challenging, and the child needs to downgrade, for instance. Verifying *if a child is challenged* might be game-specific, but the adaptation mechanisms are general to all games. Each game relies on a set of common variables such as the time to a player's action, the time to appear an object, the number of trials, and how fast objects appear. The adaptation mechanisms could inform the games on these specific parameters.

#5. Preliminary empirical results

In this section, we discuss the training of predictive models based on data from a pilot study conducted in May-June 2023. The main goal of the computational experiments is to perform an initial assessment of how well the students' game performance can predict teacher evaluations. These evaluations correspond to answers to questionnaires about multiple aspects of students filled out by the teachers *after* the students have played the games. In our situation, the teacher evaluations can be seen as a *ground truth* to the AI algorithms. Finally, we also investigate whether collected sensor data affects the results.

Data The dataset consists of all students who participated in the pilot study. The data of every student was randomised at this point; thus, we did not have any data allowing us to trace back the person from whom the data was collected.

The dataset has N=24 rows, each one representing one student. We selected the following features for each student:

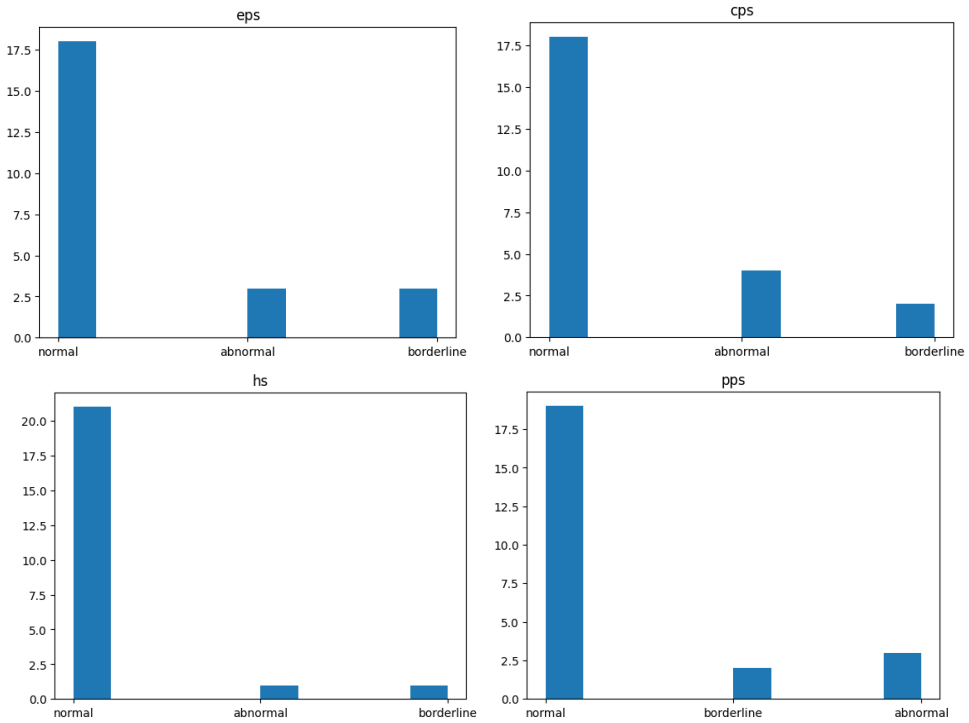
- **Individual characteristics:** *age, diagnosis*

- **Performance in games**. Each student played all three games. For this analysis, we considered only the performance on each game's first level of difficulty.
  - Attention game: *Totals of correct present, correct absent, error present, error absent; mean response time.*
  - Working memory game: *Totals of correct pepper classify, incorrect pepper classify, correct sorting, correct not sorting, long sequence.*
  - Inhibition game: *Totals of error absent, correct present, correct absent; mean responsive time, mean responsive time present, mean responsive time absent*.

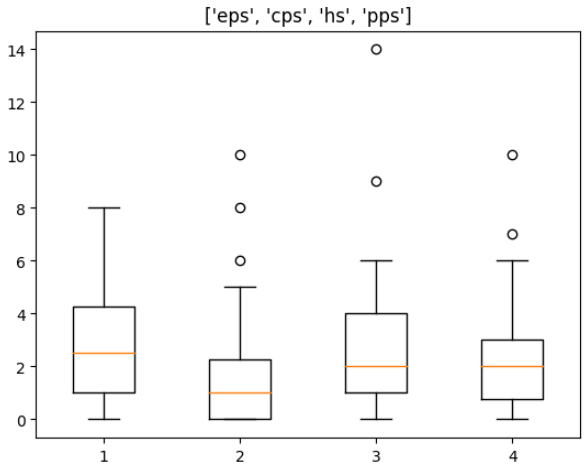
- **Sensor data**. A smartwatch collects heart rate measurements. Various factors, including bugs during the pilot limited the data availability. This impacted the collection of valid fine-grained measurements. All the features are "Root Mean Square of Successive differences between normal heartbeats" (RMSSD) at different levels:
  - RMSSD during the *attention game*, *working memory game, inhibition game*, and *no activity* (the period after the student has played all the games)
  - *Every 5 min* (*13 rounds*).

- **Teacher evaluation**: Strength and Difficulties Questionnaire (SDQ), filled out by the teachers. Based on discussions with the psychologists from WP2, SDQ subscales were used since they were indicated to be more relevant than the total SDQ alone. Each subscale is a continuous variable with a range 0-10, but they can alternatively be taken as a categorical variable, with bands indicated next to each scale:
  - *Emotional Problems Score (EPS):* 0-4, 5, 6-10 (normal, borderline, abnormal)
  - *Conduct Problems Score (CPS):* 0-2, 3, 4-10
  - *Hyperactivity Score (HS)*: 0-5, 6, 7-10
  - *Peer Problems Score (PPS)*: 0-3, 4, 5-10

Descriptive statistics for each SDQ subscale, a dataset was built with the same predictors and the corresponding SDQ subscale as the class variable to be predicted. The figure below shows the class distribution for each subscale.



The figure below shows the distribution of the SDQ subscales in the original (numerical) form.



Methods

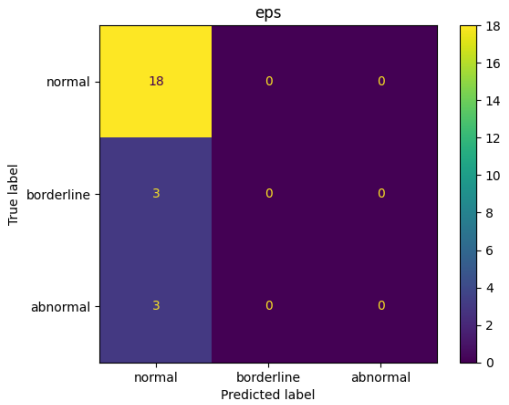
Data pipeline: we designed a data pipeline that performs data cleaning according to the dataset's characteristics, which is described as follows.

- **Imputation:** imputes missing values in each feature using the strategy of most frequent matters in the feature.
- **Encoding of categorical features:** categorical features are encoded into numerical features using the one-hot encoding scheme.
- **Scaling:** once the dataset is transformed, it has only numerical features, which are scaled by computing normalised scores: (x - m(f))/sd(f), where x is a value of feature f, m and sd are the mean and standard deviation of the feature f, respectively.

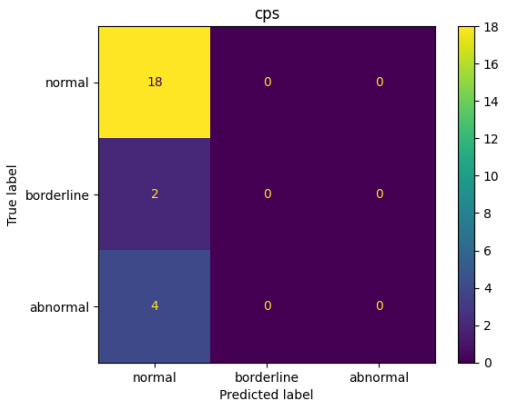
After pre-processing the data, we trained a Random Forest Classifier to predict the subscale score. We used Python and the library scikit-learn as the backend library for machine learning.

For model evaluation, a 5-fold stratified cross-validation was used.

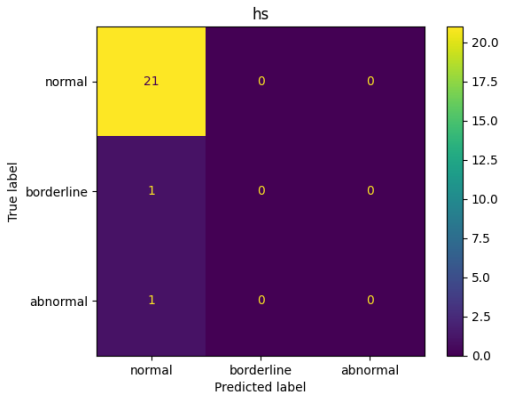
Results The figure below shows the confusion matrices for predicting each SDQ subscale *without* using the sensory data.



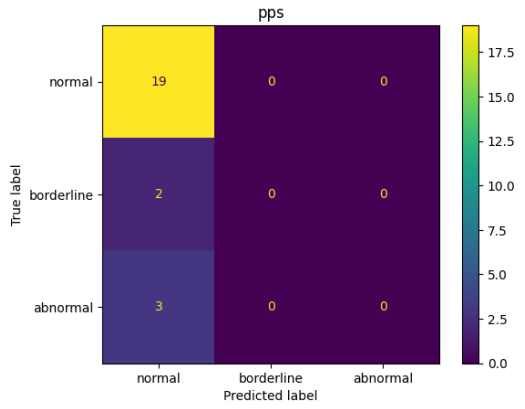
Acc = 0.75; F1 (macro) = 0.29



Acc = 0.75; F1 (macro) = 0.29



Acc = 0.91; F1 (macro) = 0.32



Acc = 0.79; F1 (macro) = 0.29

The results show that the trained ML models consistently predicted the majority class ("Normal" label). We did not observe any changes in the results by including the sensory data (heart-rate variability).

The average-quality predictions indicate that the models need further improvement. The main reason for such model performance is likely the very small dataset (N=24 subjects). In addition, the available features on game performance and sensory data are likely not strongly related to the predicted outcomes (SDQ subscales). This is likely due to the lack of fine granularity in-game performance and sensory data.

We expect that the following pilot study will be able to use a more complete data collection procedure to increase the predictive power of the algorithms.