# E-Health Methods & Applications: Project report Part IIA GROUP 3

## 1. Introduction

The aim of this project is to identify the relevant serious games specific for ADHD (Attention Deficit Hyperactivity Disorder) and LSDs (Learning Specific Disorders) present on Google Play Store that are scientifically validated, and provide for those a characterization summarizing their main information. The idea is to create, then, a useful interface representation of the mentioned above market, for companies that want to develop a new application in this field.

## 2. Materials and methods

### 2.1 Apps selection

The starting point is the previously created database during the first stage of this project, consisting in a collection of serious games for children available in Google Play Store: this was obtained by means of the development of an automatic algorithm in Python which filters out the serious games based on the category and the description of the app, to select only the possible candidates. It is necessary to state that with respect to delivered database at the end of stage 1, integrations of additional apps have been done by further retrieving the app ID in case of incorrect-ID error. Thus, through a previously created Python function, using *google search* library to query for the app name followed by the words “google play”, the app ID is retrieved when possible and information from it are extracted. The result of this further step is the broadening of the database from 5630 to 5877 apps (considering also the elimination of apps having “contentRating” excluding children from users: “Teen”, “Mature 17+”, “Everyone 10+”).

Therefore, the serious games belonging to ADHD or LSDs fields were selected. In particular, the selection of the apps of interest was performed by automatically searching in the title or in the summary of the app (features stored in our database) the word-roots “dysgraph-”, “dyslex-”, “dyscalcul-”, “anorthograph-”, “dysprax-” for LSDs field and the word “ADHD” for ADHD field. The discarded apps from this selection are still reported in the dashboard for completeness. To simplify following steps and to speed up calculations, the apps of interest and the ones excluded were stored into two different tinyDB databases.

### 2.2 GoogleScholar scraping

As mentioned before, the idea to reach the goal of the project is to find out if the selected apps are discussed in scientific papers and, thus, have scientific evidence. To do this, the implementation of an automated generation of the url corresponding to the page of results obtained by the search on Google Scholar of the app name followed by the word “app” (using *urllib* library) is done.

The url generated is then used by the Python *Request* module to fetch the data, subsequently *BeautifulSoup* is exploited to extract information and to use CSS (Cascading Style Sheets) selectors to query the page for meaningful data. In the search result page, each of the HTML item is encapsulated in a tag with the attribute data-lid having null value. This is used to break the HTML document into data-lid elements containing information about individual items. From that, the urls of the results in the first search page of Google Scholar are extracted which will be exploited to access to papers content.

## 2.3 Abstract extraction

In order to obtain the needed information for the characterization of the serious games, the abstract of the resulting paper is explored. Considering that papers are stored in different websites of scientific literature, each one having its own page template and so its own html structure, the choice performed was to consider 6 of the most popular ones (Springer, Nature, Eric, IEEE, Science Direct, Scopus) and create site-specific python functions to extract the abstract of the paper by means of HTML scraping. Thus, these functions are applied to the urls returned by the search on Google Scholar that match the base-form url of the specific site (comparison done using *re* library).

## 2.4 Natural Language Processing

To simplify the further analysis of the extracted abstract, a Natural Language Processing NLP function was created with the aim to normalize and standardize the text. Firstly, the subdivision of the text in substrings corresponding to single words (tokenization) was computed. Thereafter, the words were classified as their corresponding part of speech (POS-tagging): this was done in order to allow a correct lemmatization (reduction to base form). So, at the end, the NLP function returns a dictionary containing words (without duplicates) in their base form. All the steps were done by the usage of *nltk* library.

## 2.5 Classification of study type

The next step comprises the assessment of the level of clinical evidence produced on selected serious games, depending on the type of study carried on in associated papers found. This is performed using provided lists of words characteristic for each study type, which sorted by reliability are: observational study, systematic review, randomised control trial and meta-analysis. Since the provided dictionaries contain words that are not in their base form, the function Natural Language Processing was applied to them. The classification of the selected serious game papers into the different study types is a multi-variate classification problem. This was carried on by counting the number of words in common between the abstracts (after NLP processing) and the four dictionaries. Thus, the paper is classified with the study type whose correspondent dictionary shows the highest number of words in common with the abstract. Whenever the highest number of words in common is shared among more than one study-type dictionary, it has been chosen to classify the paper with the study type with the major reliability. Moreover, if the number of common words is zero for each study type, then it is classified with the label “other”.

## 2.6 CSV file creation

After the collection of information related to the papers, other details belonging to the applications were added. In particular, due to the fact that in the database created during the first part of the project some of wanted information weren't stored, a function was created in order to retrieve them. In this function both *googlesearch* and *google play scraper* libraries were employed with the aim to access to information of the app's release, the scores given by users and the app category.

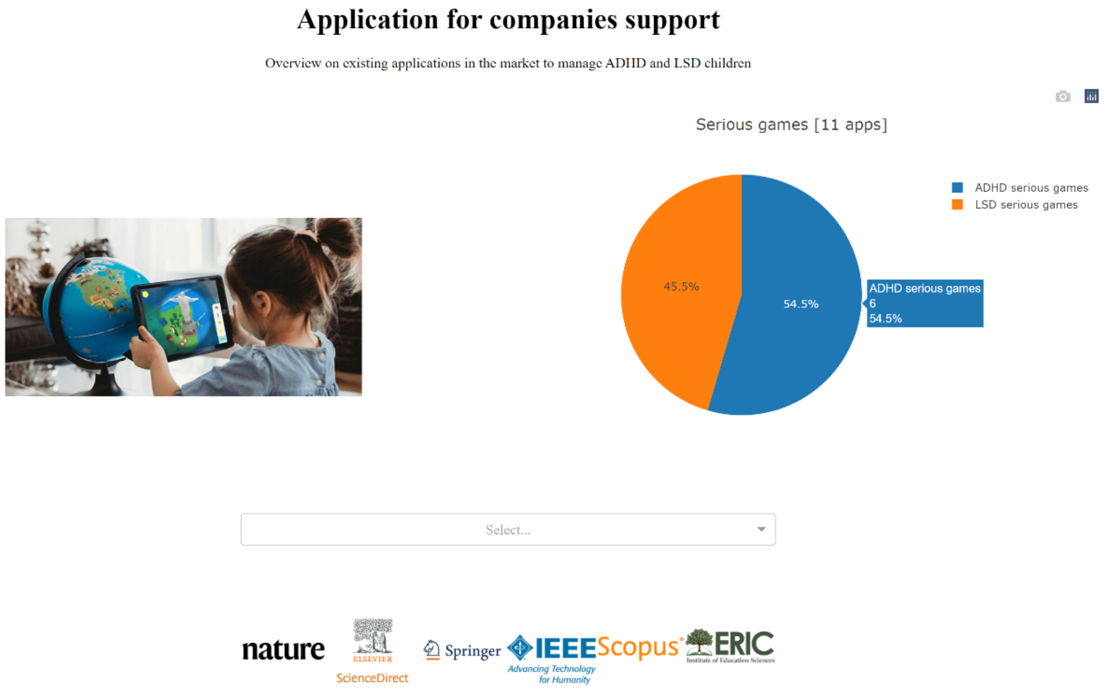
At the end two csv files were created (one for applications related to ADHD, the other one to LSDs) specifying for each application the app name, the developer, the release, the score given by users, the price, the paper's urls (all those resulting in the Google Scholar page that come from one of the scientific literature web sites treated) and the study type of the paper for each collected url.

In addition, another csv file was created containing only the apps name and the category for those apps which didn't belong to the selected fields.

## 2.7 Dashboard

To better visualize the results, a fairly simple and user-friendly graphical interface was created by the usage of *dash* library. It consists in a web page where the user can access to information just extracted. This was done with the view to provide to companies that want to develop a new application for ADHD and LSD children data related to serious games already present in the market.

First of all, a pie graph shows the number of applications related to the two chosen fields (ADHD and LSDs). Then, by a dropdown menu the user can select the field of apps he wants to inspect:  ADHD serious games, LSDs ones or other serious games not related to the previous fields. After the choice, it is possible to access to some data: for the fields of interest, it is present a table which shows on columns the app name, the developer, the first release, the score given by users, the price, the urls of papers specific for each app and the characterization of the papers.



Application for companies support

Overview on existing applications in the market to manage ADHD and LSD children

*Figure 1: Dashboard*

Regarding the serious games not belonging to the chosen fields, a table containing only the app name and the app category can be visualized.

Concerning the data shown in the different tables, they are retrieved from three csv files (one for each field). In addition, to better visualize the results, only ten applications belonging to ‘other’ field are shown.

Lastly, also some images are present in the dashboard: at the beginning of the page is present a picture showing a child which is playing with a tablet, while at the end of the web page it is possible to see the logos of the websites where papers are stored.

## 3. Results

Among the total amount of apps in the original database of 5877 apps, the ones relative to the field of interest turn out to be 11 (5 for LSDs and 6 for ADHD).

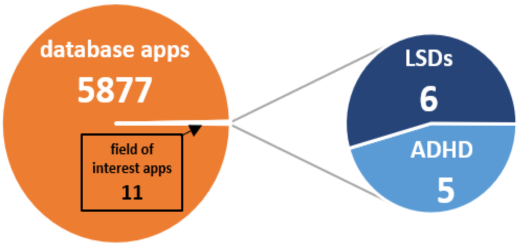
With Google Scholar search, for 10 apps among the ones of interest there is at least one paper url whose abstract is accessible by means of the created website-specific functions.

Concerning the pertinence of the results coming from the search on Google Scholar with respect to the apps of interest, it is appreciable whenever the app name is not generical, yet specific and unique in identifying the distinct app. For this reason, it happens that for some of the selected apps there is no correspondence between the app and the paper content. In particular, it is found out that 6 above the set of 22 abstracts explored (one or more for each app) contain information content relative to the corresponding app.

Finally, regarding the classification of the study type, the choice of classifying the study type of a paper with the one with major reliability in case of equality between two types, is dictated by the hierarchical structure of the classification of medical literature, where higher reliability level incorporates lower levels. The performance of such classifier (computed considering all the abstract retrieved and not just the ones actually inherent to the corresponding app) turned out to be 15 over 22 papers.

| **Parameter** | **Value** |
|---|---|
| Apps in the field of interest | 11 |
| Apps relative to LSDs | 5 |
| Apps relative to ADHD | 6 |
| Apps having at least one accessible paper | 10 |
| Total number of papers accessed | 22 |
| Papers inherent to the relative app | 6 |
| Papers with correct study-type classification | 15 |

*Table 1: Numerical values of results.*



*Graph 1: Pie chart of database subpartition.*

## 4. Limits

During the development of this project, different problems regarding mainly limited time and scraping came out. The major ones faced during the implementation of this second part of the work were:

- Google policy against web scraping: limiting the frequency of requests for urls and for getting access to the paper abstract becomes a real problem considering a possible future scale enlargement of the project (issue not encountered in our case because dealing with a relatively small number of apps, specifically related to ADHD and LSDs field);
- Retrieval of papers which often are not referred to the searched app by Google Scholar, whose possible reason can be the excessive generality or commonality of the app name;
- Creation of abstract-extraction functions just for a limited number of scientific literature websites due to time constraints.

## 5. Conclusion

In conclusion, our work has led to the identification of serious games intended for children already present in Google Play Store with a particular focus on those related to ADH and LS disorders. This latter activity was implemented with the aim to create a tool for companies that want to develop a new serious game specific for this field. In addition, specific publications were retrieved in order to see if the found apps were scientifically validated and in case of the presence of publications, the classification of the study type was done, achieving good results in it (particularly in case of inherence between article and app).

Regarding the possible future developments, they could include an automatic updating of both the app database and the articles, and for instance, this could be accessed by inserting an "update" button in the dashboard. Another improvement could be the extension of the field of interest, by considering apps relative to a wider set of disorders (for example autism and motorial disorders). Lastly, concerning the scientific literature websites, additional site-specific python functions for abstract extraction could be implemented in order to allow a more complete view on the app validation.