ML4MICROBIOME WORK PACKAGE 1

State-of-the-art evaluation and update

Overview of activities

OBJECTIVES

To continuously evaluate the state-of-the-art ML/statistics methods, and to ensure that every action member is "on the same page" in terms of their robustness and suitability for microbiome research and how well they address the specific challenges in 1.1.1, separately and combined.

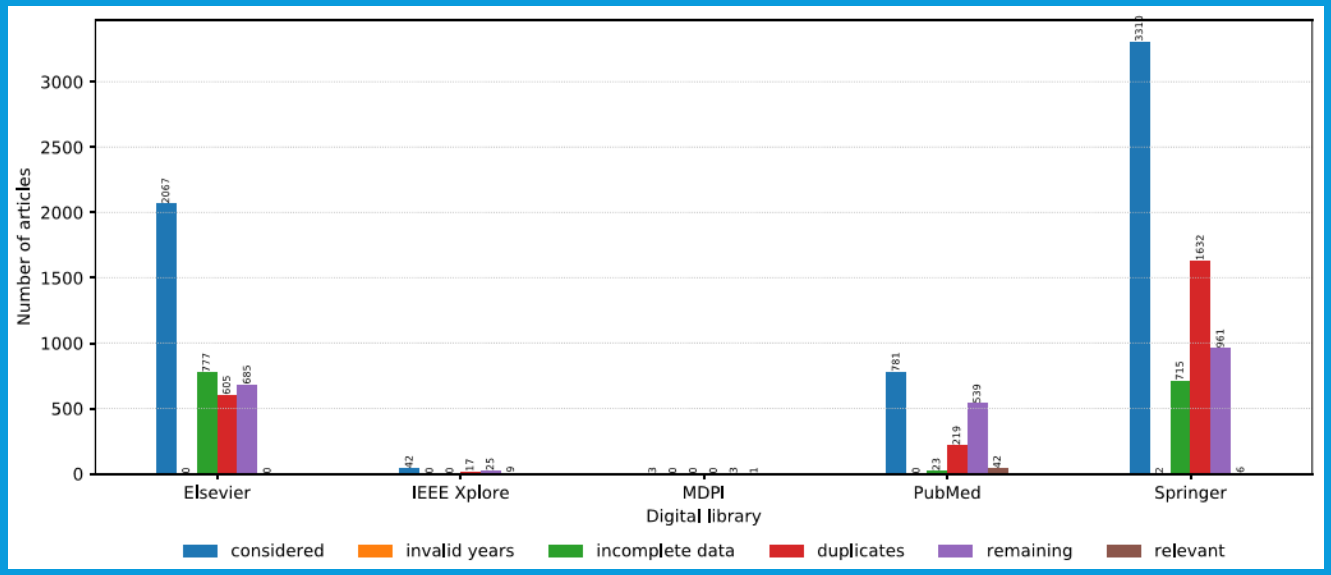
Task 1.1: Technology watch

Task 1.2: Evaluation of ML/statistics methods currently used in microbiome research

Task 1.3: Define priority areas for novel ML/statistics applications for microbiome data

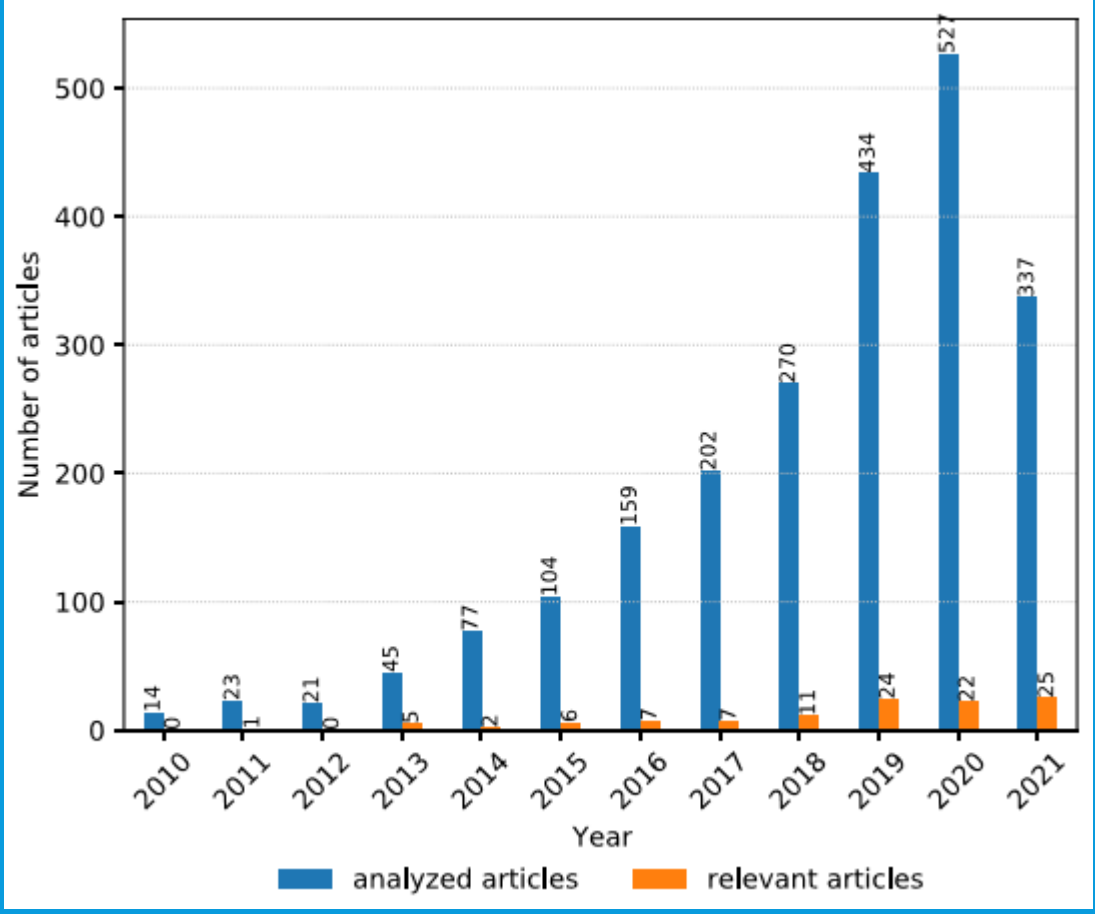
SCOPING REVIEW

- In 2020, Springer, IEEE, and PubMed were accessed
- In 2021, two additional digital libraries: Elsevier and MDPI were included
- No Oxford Academic Publishers included



NLP toolkit (Zdravevski et al., 2019)

PUBLICATION SEARCH RESULTS



MICROBIOME RESEARCH IN THE SOURCE CODE REPOSITORIES

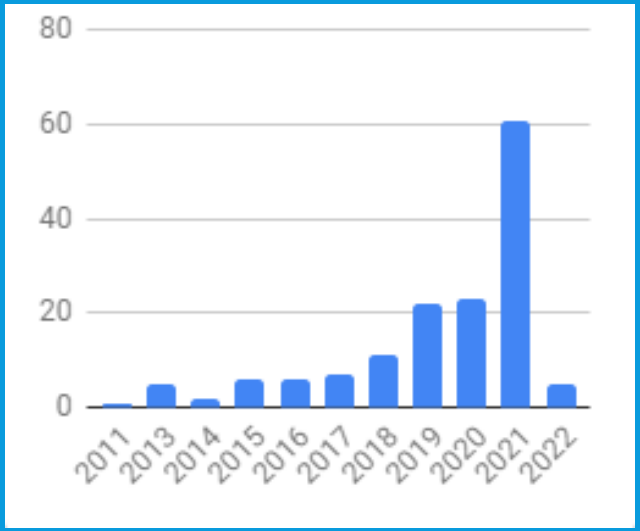
Short summary (from September 2021)

- Number of repositories 1855 (old), 2465 (new).
- New repositories 1014.
- Repositories deleted 404.

- Update (October)
  - As of today we have slightly more repositories (2589).
- Updated info on github list http://microbiome.przymus.org/
- Next step – identify new (unique) publications for 2021
- Number publications for year 2021 – 4 (Dec 2021)

PERFORMED TASKS

1. Data base curation – Scoping review
2. Data base creation/curation – Github
3. Additional search (OAP)
4. Merging the obtained data bases
5. Updating the data base by action members (Feb-March 2022)
6. Report compilation (June-Aug 2022)



ARTICLE DATABASE GENERATION

- An automated search of digital libraries of three major publishers (PubMed, Springer, Elsevier, MDPI and IEEE) using NLP Toolkit (Zdravevski et al., 2019) to automate the literature search, scanning, and eligibility assessment. This method yielded **25 papers**.

- An automated search through the available GitHub resources using NLP algorithms to identify relevant software repositories and extract corresponding scientific papers. The papers were automatically ranked by relevance using the pointwise learning to rank approach (Fejzer et al. unpublished) trained using the manually collected and labeled papers. The final list includes **four papers**.

- Manual search – crowdsourcing of the studies relevant for the review topic by all members of the COST Action CA18131 "Statistical and machine learning techniques in human microbiome studies". In this way, **33 papers** were added to the final list.

After revision - **41papers** in final list.

ANALYSIS OF COLLECTED ARTICLE DATA SET

- 41 papers were adopted for the year 2021, indicating a substantial increase in the application of ML methods for human microbiome analysis compared to previous years.
- The primary disease type in collected articles was inflammatory bowel disease (19%), followed by drug-related side effects and adverse reactions, and diabetes.
- More than 70% of studies have used amplicon sequencing data (16S rDNA) and 10% only shotgun metagenome data as input data type.
- The most often used methods were random forest, logistic regression and support vector machine. Comparison with an earlier data set shows that random forest is still the most used method, but the application of logistic regression and support vector machine algorithms has increased.

ML SOFTWARE MANUSCRIPT

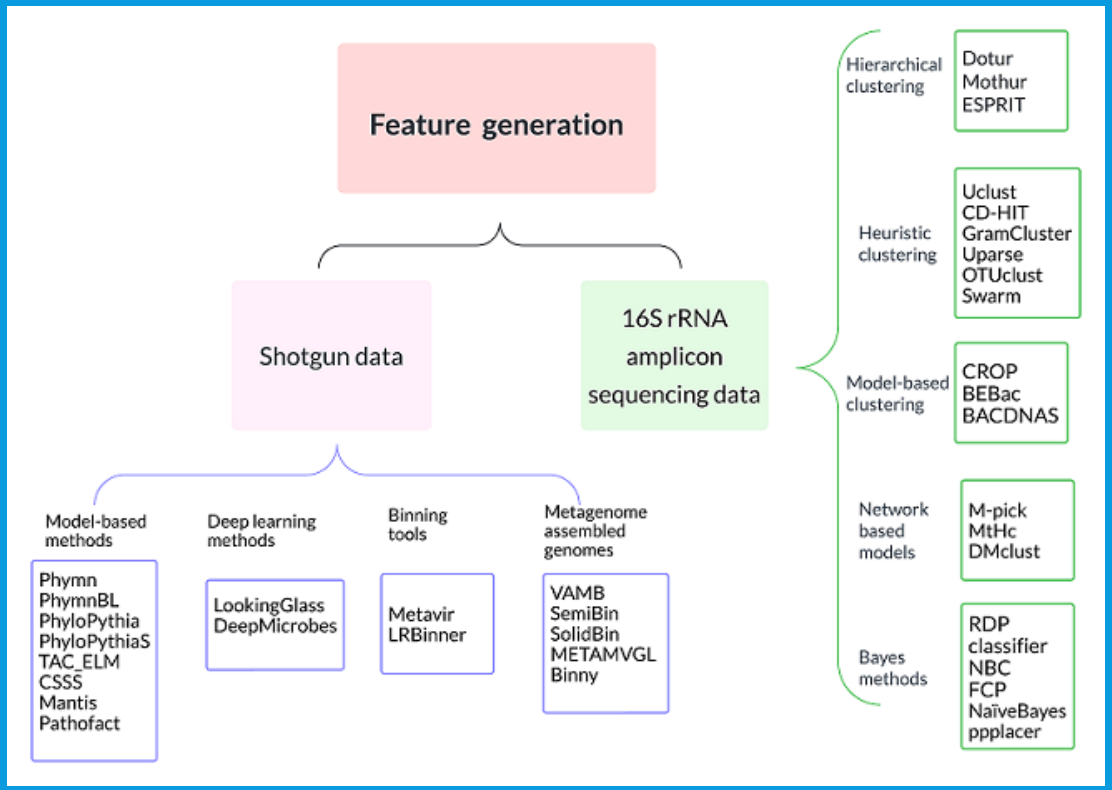
MICROBIOME

- The aim is to go beyond the application of ML techniques in the microbiome field and focus on the review of **ML-based software** and framework resources currently available for the analysis of **microbiome data in humans**.

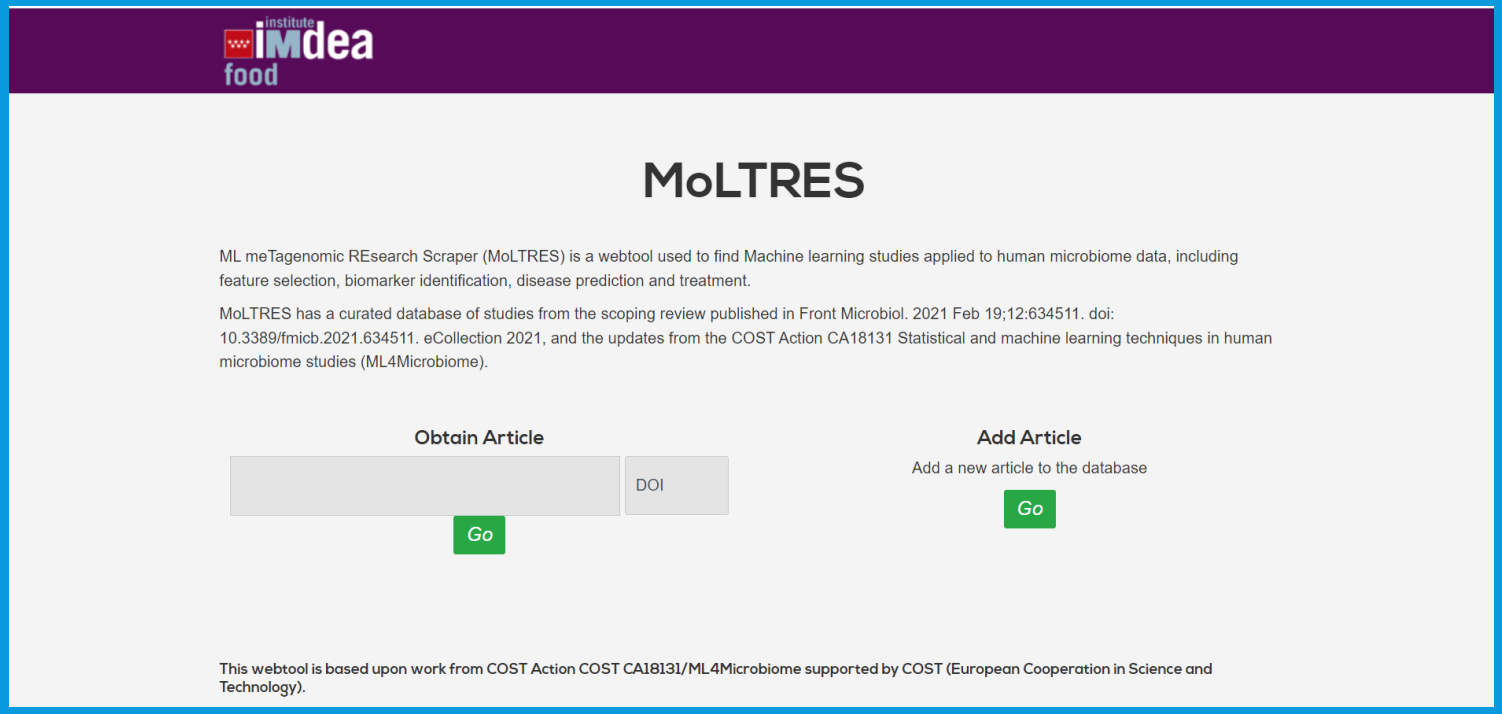
- Provide:
  - A description of each software with examples of usage.
  - Comments about pitfalls and lacks in the application of ML methods in relation with microbiome data that need to be considered in software development.

ML SOFTWARE MANUSCRIPT

MICROBIOME



WEB-TOOL FOR RETRIEVING ML HUMAN-MICROBIOME RELATED STUDIES REVIEWED/SELECTED BY THE COST



REPORTS FOR YEAR 2023

- Two reports for this final year were listed:
  - A results of publication review for the year 2022.
  - Report outlining the trends and outlook for the application of ML in microbiome data analysis.
- Replace a report outlining the trends and outlook with a section in an opinion paper prepared by WG4.

WORK STILL IN PROGRESS

- List of publications for year 2022
- ML software article
- Next paper draft about deep learning methods in microbiome in microbiome data analysis