FINDING BIOLOGICAL MARKERS FOR THE PREDICTION OF COLORECTAL CANCER

References

1] Vega et al. (2015). Colorectal cancer diagnosis: Pitfalls and opportunities. World J. Gastrointest. Oncol., 7(12), 422–433.

[2] Pasolli et al. (2017). Accessible, curated metagenomic data through ExperimentHub. Nat. Methods, 14(11), 1023–102.

[3] Allali et al. (2018). Gut microbiome of moroccan colorectal cancer patients. Med. Microbiol. Immunol., 207(3- 4), 211–225.

[4] Wirbel et al. (2019). Meta-analysis of fecal metagenomes reveals global microbial signatures that are specific for colorectal cancer. Nat. Med., 25(4), 679–689.

01 INTRODUCTION

- **Current colorectal cancer** (**CRC**) **detection methods are challenging**, or techniques are used to predict the disease based on already present symptoms, so not in the earliest stage of the disease [1].
- This research aims to verify and discover **functional biological markers** for CRC using machine learning methods on a **metagenomic dataset**.
  - **Functional biological markers**: specific genetic features associated with particular functional traits or activities in the microbiome.
  - **Metagenomic data set**: non-host DNA from the human gut.

02 RESEARCH QUESTION

Can **neural networks** with wrapper methods for feature selection be used to analyse a **metagenomic data set** to verify **functional biological markers** for the disease colorectal cancer?

1. Use a **logistic regression model** to classify diagnosed and healthy samples.
2. Use a **neural network model** to classify diagnosed and healthy samples.
3. Evaluate **metrics** (Confusion matrix, accuracy, precision, recall, F1 score) for both models.
4. Identify important features by the feature selection and check if they correspond to **biomarkers** in literature.

03 METHOD

Data

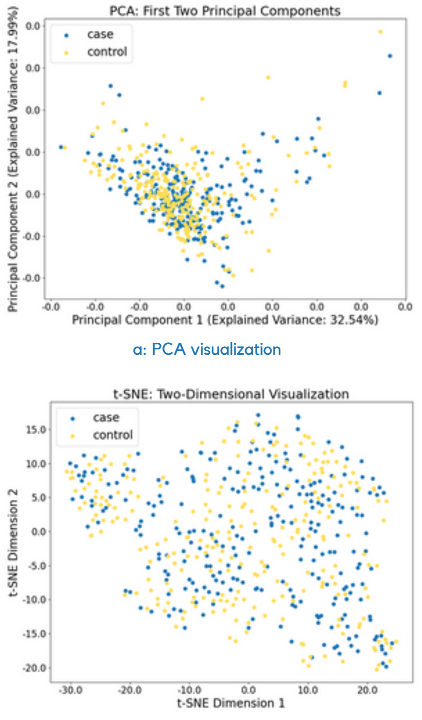
- Fecal shotgun metagenomic study of CRC from CuratedMetagenomicData [2] of which **509 samples** are used (CRC and control groups).
- **Pathway abundance** data is used, which is functional data about series of interconnected biochemical reactions or processes. This is done because multiple papers suggest this kind of data is useful for predicting CRC [3, 4].

Feature Selection and Model

- **Feature selection** techniques used to identify potential biomarkers are Recursive Feature Elimination (RFE), Forward Feature Selection (FFS), variance filtering, minimum Redundancy Maximum Relevance (mRMR), and Principal Component Analysis (PCA).
- **Machine learning models** to evaluate the selected features and assess their predictive power are logistic regression and neural network (multi layer perceptron).
- Features with high importance are identified as potential **biomarkers.**
- **Feature importance** is defined as absolute coefficients for logistic regression.
- For the neural network permutation importance is used as feature importance.

04 RESULTS

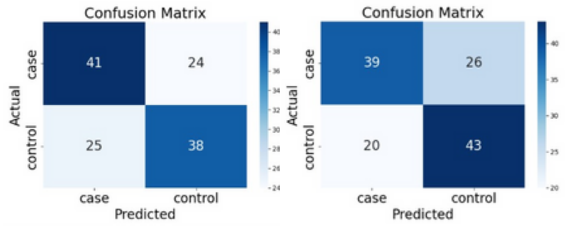
- Case and control datapoints show a **lot of overlap** when plotting them using PCA and t-SNE. This can be seen in Figure 1.
- Logistic regression and neural network **score similarly low on accuracy** with and without feature selection.
  - Feature selection by variance gets the best average cross validation accuracy scores: **0.58** for logistic regression and **0.57** for the neural network. Confusion matrices from runs on the test partitions are visible in Figure 2.
- In Figure 3 can be seen that out of the top 10 most important features from both models **4 pathway abundances overlap with literature**.
- 3 pathways overlap between logistic regression and the neural network, which are **potential biomarkers** (figure 3):
  - the superpathway of pyridoxal 5′-phosphate biosynthesis and salvage;
  - the superpathway of mycolate biosynthesis;
  - peptidoblycan biosynthesis V.



a: PCA visualization

b: t-SNE visualization

Figure 1: The data visualised with PCA and t-SNE.



a: Logistic regression

b: Neural network

Figure 2: The confusion matrix of the models combined with feature selection by variance filtering ran on a test partition.

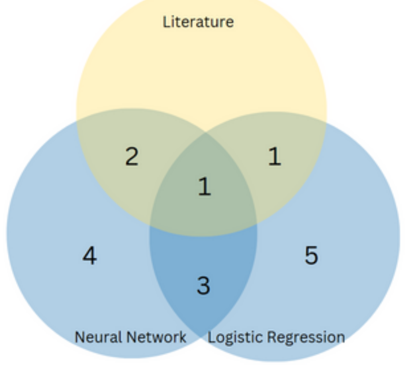


Figure 3: Venn Diagram of most important features selected by variance filtering. Numbers indicate the amount of pathways.

05 CONCLUSION & LIMITATIONS

Conclusions:

1. Logistic regression and neural network models have **similar performance.**
2. **Overlap** between features with high importance and literature is found, making the other important features interesting for further research.

Limitations:

- The achieved accuracy levels may be too low to draw definitive conclusions about the **relevance of the selected features**.
- Used methods only provide a **first indication** of potential biomarkers for CRC. Further research is necessary to confirm their relevance.

Author: Jos Sloof (A.J.G.Sloof@student.tudelft.nl, jsloof)
Supervisors: dr. Thomas Abeel, David Calderón Franco1, Eric van der Toorn

TUDelft