DETECTING SCHIZOPHRENIA WITH MACHINE LEARNING

Using Machine Learning to identify Schizophrenia-related Biomarkers with data derived from the gut microbiome

1. INTRODUCTION

Diagnosis of Schizophrenia relies on psychological assessment of the patient leading to identification of the disorder only after it has reached an advanced stage [1]. There is mounting evidence that mental disorders affect **the relation between the gut microbiome and the brain** [2].

Only a handful of studies have attempted to utilize Machine Learning to find schizophrenia-related biomarkers using data from the gut microbiome.The research presented here aims at **verifying biomarkers** found in the literature.

2. RESEARCH QUESTIONS

1. **What Machine Learning models could be used on meteganomic data from the gut?**
2. **What are the relevant biomarkers and do they correspond with findings from literature?**
3. **How reliable are the models at predicting Schizophrenia?**

5. RESULTS

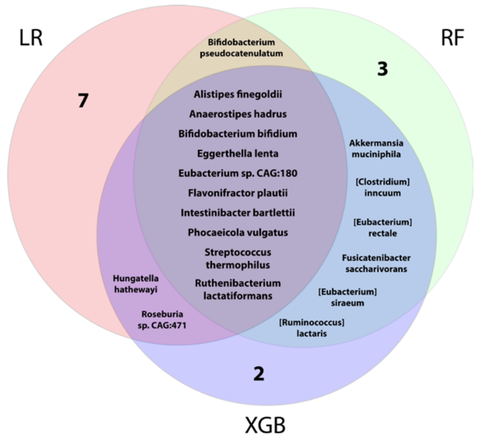

Figure 1: Venn diagram showing the overlap between each classifier's top 20 most important features across all cross-validation runs (before predictions on the validation set). There are 19 overlapping species out of 31 distinct species present.

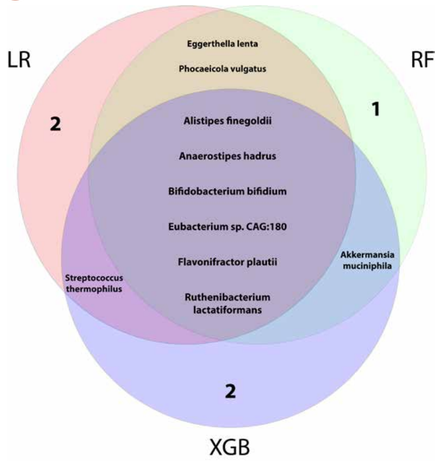

Figure 2: Venn diagram showing the overlap between each classifier's top 10 most important features ranking across predictions made on the validation set. There are 10 overlapping species out of 15 distinct species present.

Three classifiers - **Logistic Regression** (LR), **Random Forests** (RF), and **XGBoost** (XGB) - were used to extract feature importance. Each classifier underwent a cross-validation process and feature importance was calculated. Overlapping features between the top 20 most important features of each classifier were retained. Each classifier then used this new set of features to make predictions on a validation set. This process was repeated 10 times and the importance of features across the iterations was averaged.

The performance of each model during both the cross-validations runs and the predictions on the validation set are shown in Table 1 and Table 2 respectively. Performance is similar for RF and XGB but **LR** shows an **increase in performance** after **updating the feature set** .

There is a total of 19 overalpping features out of 31 distinct features present in the classifiers' rankings. RF and XGB, which have similar performance, agree on the importance of a high number of features during the cross-validation process, in contrast to LR. After updating the feature set, the three classifiers have

Table 1: Table showing the average performance metrics of the three classifiers across the cross-validation runs (**before** only overlapping features are used for training). RF and XGB have similar performances whilst LR considerably **underperforms**.

	LR	RF	XGB
Accuracy	0.54	0.67	0.67
Precision	0.58	0.69	0.71
AUC-ROC	0.57	0.74	0.74

Table 2: Table showing the average performance metrics of the three classifiers across the predictions on the validation set (**after** only overlapping features are used for training). RF and XGB have similar performances whilst LR considerably **overperforms**.

	LR	RF	XGB
Accuracy	0.73	0.63	0.64
Precision	0.7	0.62	0.62
AUC-ROC	0.73	0.62	0.64

3. DATA

171 samples (90 schizophrenia, 81 controls) were obtained through the CuratedMicrobiomeData package made available by [3]. Gender, age, gender, and orgin are well distributed among the cohorts. The data was processed through **Shotgun Sequencing**, a DNA sequencing method that is reliable at obtaining **species-level** data from strands. The **taxonomic relative abundance of species** was extracted from the data and used to train the models.

6. CONCLUSION

Eubacterium sp. CAG:180 and *Ruthenibacterium lactatiformans* emerge consistently as the most important species but are not reported in the literature [3][4]. 8 other specie rank as important across the three classifiers and are reported in the literature [3][4].

Logistic Regression underperformed during cross-validation but excelled with the validation set, possibly due to the selection of features by the other classifiers. Further research is recommended to examine the importance of identified species, enhance statistical testing, and expand to genus-level analysis for broader comparability.

REFERENCES

[1] Lee, J. Seo, H. C. Jeong, H. Lee, and S. B. Lee, "The Perspectives of Early Diagnosis of Schizophrenia Through the Detection of Epigenomics-Based Biomarkers in iPSC-Derived Neurons," in Frontiers in Molecular Neuroscience, vol. 14, 2021. https://www.frontiersin.org/articles/10.3389/fnmol.2021.756613

[2]D. Wang, W. A. Russel, Y. Sun, K. D. Belanger, and A. Ay, "Machine learning and network analysis of the gut microbiome from patients with schizophrenia and non-psychiatric subject controls reveal behavioral risk factors and bacterial interactions," Schizophrenia Research, vol. 251, pp. 49–58, 2023, doi: 10.1016/j.schres.2022.12.015.

[3] F. Zhu et al., "Metagenome-wide association of gut microbiome features for schizophrenia," in Nature Communications, vol. 11, 1612, 2020. https://doi.org/10.1038/s41467-020-15457-9.

[4] E. Castro-Nallar, M.L. Bendall, M. Pérez-Losada, S. Sabuncyan, E.G. Severance, F.B. Dickerson, J.R. Schroeder, R.H. Yolken, and K.A. Crandall, "Composition, taxonomy and functional diversity of the oropharynx microbiome in individuals with schizophrenia and controls," PeerJ, vol. 3, 2015.

Author: Timothy Bastow t.m.bastow@student.tudelft.nl **Responsible Professor:** Thomas Abeel **Supervisors:** Eric van der Toorn, David Calderon Franco

TUDelft