Identifying biological markers in the gut microbiome associated with celiac disease using machine learning

Persianov Petr

TU Delft

1. Introduction and background

Celiac disease is a genetic autoimmune disorder that affects approximately 1.4% of the global population. The **current standard for diagnosis** is a combination of blood tests and an endoscopic biopsy of the small intestine, which are both **invasive and painful procedures**.

Recent research has suggested that the **gut microbiome**, the collection of microorganisms that live in the digestive tract, may **play a role in the development and progression of celiac disease** [2]. Specifically, **alterations in the composition of the gut microbiome** have been observed in individuals with celiac disease compared to healthy individuals [1].

Previous research has used machine learning to analyze the **gut microbiome** in various disease contexts, including inflammatory bowel disease (IBD) and colorectal cancer. These studies have shown that machine learning can be a powerful tool for **identifying disease-specific biological markers in the gut microbiome**.

2. Research questions

Main research question: Can machine learning be used to identify biological markers in the gut microbiome that are associated with celiac disease?

Sub-questions:

- Which machine learning algorithms excel in classifying celiac disease samples using gut microbiome data?
- Which feature selection methods work best with selected machine learning methods?
- What are the specific bacterial species associated with celiac disease identified by feature selection methods?

3. Methods

Dataset and preprocessing steps

1. **Raw DNA data** was collected 2 different datasets.
2. The **quality trimming** was performed using Trimmomatic tool.
3. **Taxonomic sequence classification** for all samples was performed using Kraken2 tool on **bacteria database**.
4. **Relative abundance on species level** was estimated using Bracken tool.
5. Samples were labeled with **disease states (healthy/celiac disease)** using metadata from the datasets.

After preprocessing we have **212 samples** with **4630 bacteria species as features**. Figure 1 illustrates data preprocessing pipeline.

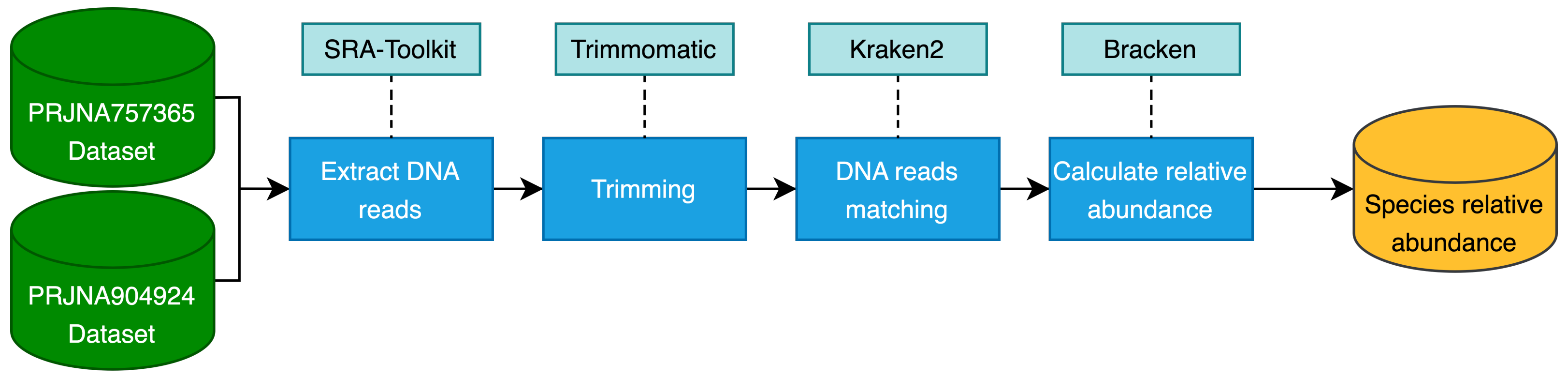


Figure 1. Data preprocessing pipeline.

Model training and evaluation

We selected the following machine learning models: **Random Forest**, **Logistic Regression**, **Support Vector Machines** and **XGBoost**. For feature selection, we chose: **ANOVA**, **Information Gain** and **MRMR**. Performance evaluation was conducted using the **Area Under the ROC Curve** metric.
Figure 2 illustrates the pipeline for training and evaluating all possible combinations of machine learning models and feature selection methods.

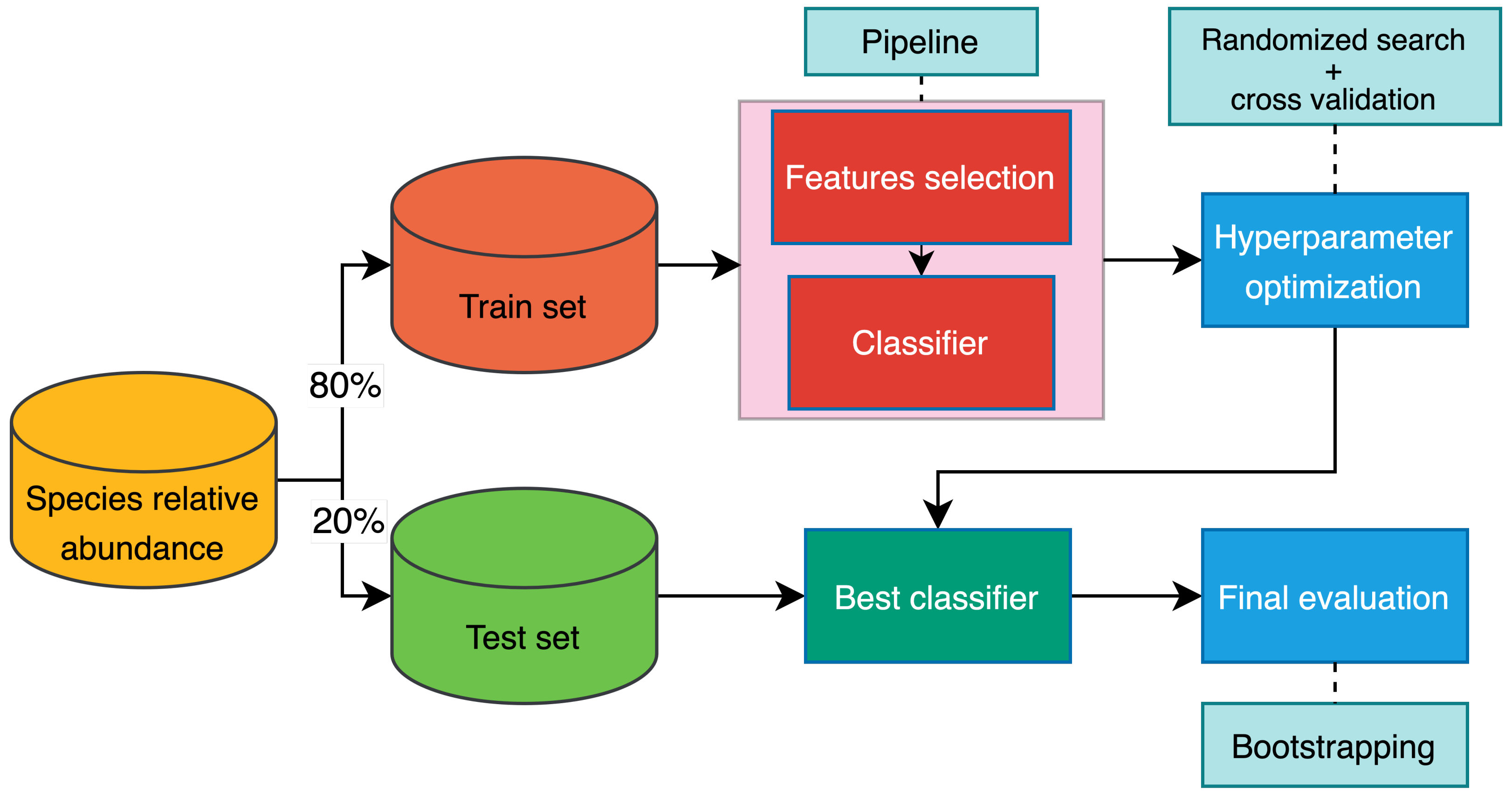


Figure 2. Model training and evaluation pipeline

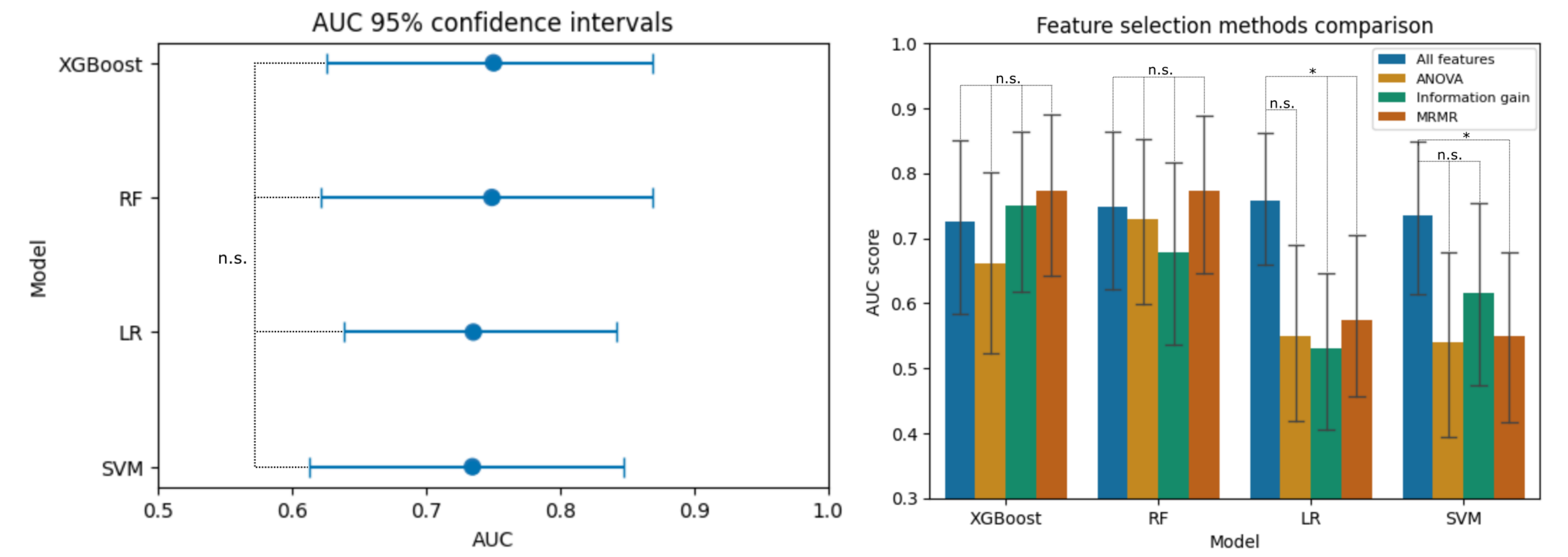
4. Results

Comparable Performance of Models without Feature Selection

No statistically significant difference was observed in the performance of models when no feature selection method was applied (Figure 3a).

Impact of Feature Selection Methods on Model Performance

Models were trained and evaluated using different feature selection methods (FSMs) as shown in Figure 2. The results, depicted in Figure 3b, demonstrate that the choice of FSM had a limited impact on the performance of the Random Forest and XGBoost models. However, for Logistic Regression (LR) and Support Vector Machines (SVM), the selected FSM significantly influenced performance.



(a) Baseline performance without feature selection.

(b) Models' performance with various feature selection methods.

Figure 3. Models' performance comparison with and without feature selection. Plots illustrate mean AUC scores with 95% confidence intervals. 'n.s.' indicates non-significant differences in performance, while '*' denotes statistically significant differences with $p < 0.05$.

Intersections between sets of selected features

11 features were commonly selected by all feature selection methods (Information Gain, MRMR, ANOVA), and their subset achieved **comparable performance to the full set of features** Figure 4.

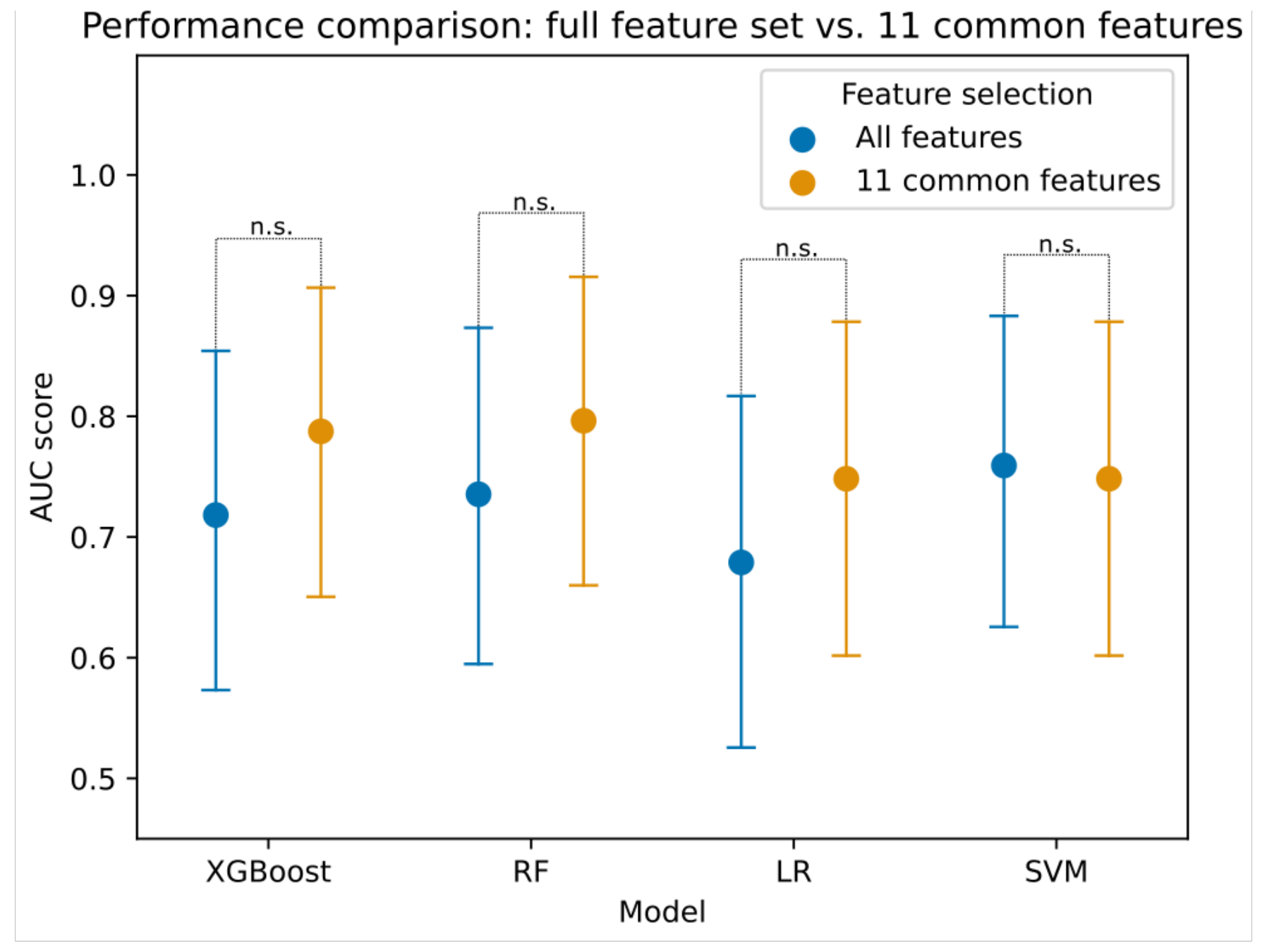


Figure 4. Models' performance comparison on full set of features versus 11 commonly selected features. 'n.s.' indicates non-significant differences in performance

Correspondence of Selected Features with Existing Findings

Bacteroides eggerthii, *Parabacteroides johnsonii*, *Faecalibacterium prausnitzii_H*, *Faecalibacterium prausnitzii_I*, and *Ruminococcus_D bicirculans* have been **previously associated with celiac disease**. Other features require further investigation as their associations with celiac disease have not been established in previous studies.

5. Conclusion

This study demonstrates the feasibility of using machine learning to identify biomarkers for celiac disease in the gut microbiome.

1. **XGBoost**, **Random Forest**, **Logistic Regression**, and Support Vector Machines showed **comparable performance without feature selection**.
2. The **impact of feature selection varied**, especially for Logistic Regression and Support Vector Machines.
3. A subset of **11 species selected by all three methods proved effective**, suggesting their potential as biomarkers. Some **selected species have been previously associated with celiac disease**, while others require further investigation.

6. Limitations and future work

This study has limitations and areas for improvement:

1. **Increasing the population size** and diversifying the sample can enhance **generalizability**.
2. **Exploring alternative feature selection methods**, such as recursive feature elimination (RFE), and **different machine learning models** can provide **further insights** into the gut microbiota's role in celiac disease.

7. References

[1] Roberta Elisa Rossi, Giulia Dispinzieri, Alessandra Elvevi, and Sara Massironi.
Interaction between gut microbiota and celiac disease: From pathogenesis to treatment.
Cells, 12(6), 2023.

[2] Lucia Sacchetti and Carmela Nardelli.
Gut microbiome investigation in celiac disease: from methods to its pathogenetic role.
Clinical Chemistry and Laboratory Medicine (CCLM), 58(3):340–349, 2020.

p.persianov@student.tudelft.nl