FINDING BIOLOGICAL MARKERS FOR TYPE 2 DIABETES

Author: Aratrika Das (A.Das-12@student.tudelft.nl)

Supervisors: Thomas Abeel, Eric van der Toorn, David Calderón Franco

TU Delft

INTRODUCTION

The human gut microbiome refers to the community of microorganisms present in our gut and has been shown to have an interdependence with disease like Type 2 Diabetes. Previous studies conducted by Qin et al and Karlsson et al observed a reduction in the phylum *Firmicutes* and *Clostridia* class in the group with T2D, along with increase in *Clostridium clostridioforme*, *Lactobacillus* and a decrease in *Roseburia*.

THE RESEARCH QUESTION

Can we use different machine learning techniques on metagenomic shotgun sequenced data of samples affected with Type 2 Diabetes and control samples to effectively identify biomarkers that can be used to predict Type 2 Diabetes ?

METHODS

EXTRACT DATA

Relative Abundance data at the **Species level** from 2 datasets - 366 Chinese samples and 145 European samples. The datasets were also combined

DATA ANALYSIS

Performed **PCA** and **t-SNE**. Filtered out features where more than 95% samples had 0 relative abundance. Data was scaled using **Z score scaling.**

FEATURE SELECTION

Maximum relevance minimum redundancy (**MRMR**) feature selection method and conditional mutual information maximisation (**CMIM**) were compared..

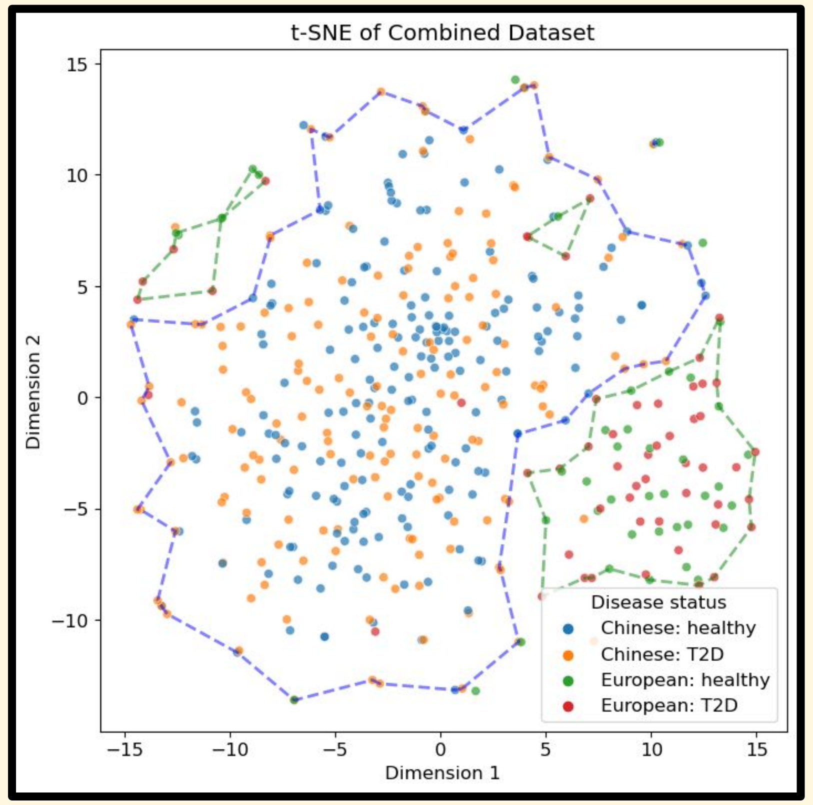
CLASSIFICATION

A **Random Forest**, **SVM, Logistic Regression, XGBoost**, classifier were trained and tested. **The accuracy, AUC, AUPRC, F1-score** were computed

BIOMARKER IDENTIFICATION

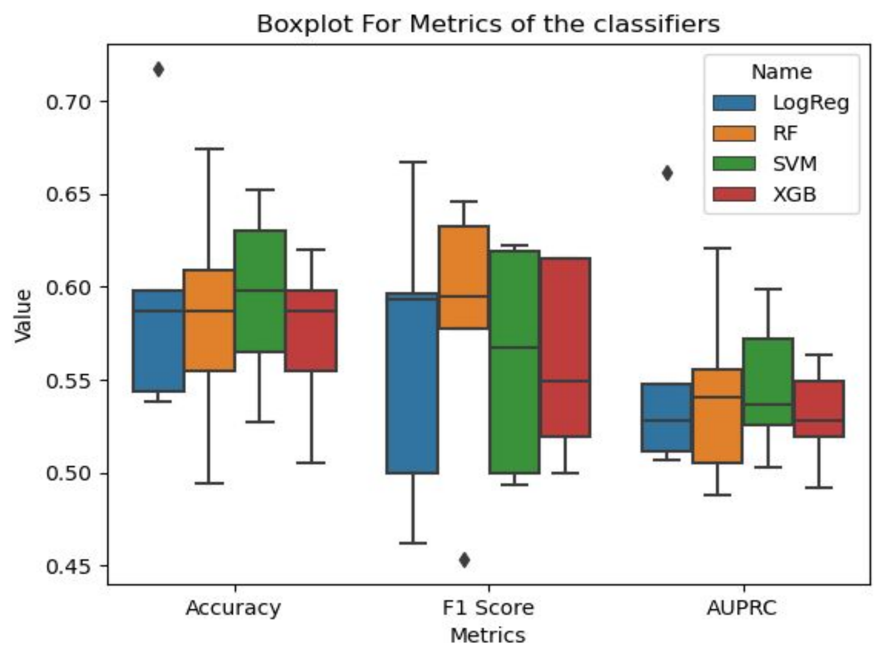
Based on the importance of the features of the classifier, the biomarkers were selected and verified with existing literature

RESULTS AND DISCUSSION

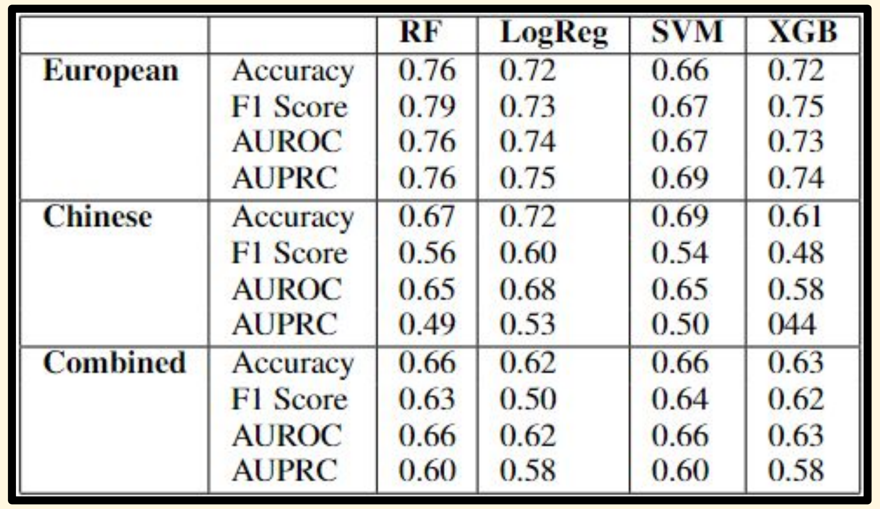


1 t-SNE on the combined data shows that data points cluster based on location

2. RF model using **mRMR performed better compared to CMIM** for all the datasets in terms of Accuracy, F1 Score, AUROC and AUPRC



3. The boxplot shows the range for the metrics for **RF, LR and SVM and XGB** models for the combined dataset during 5-fold CV. There is a large overlap in the ranges.

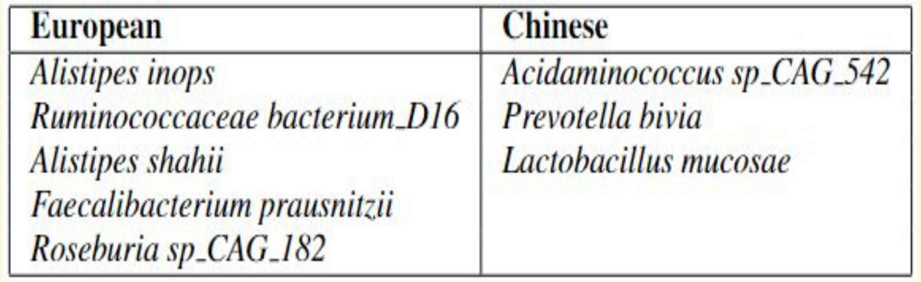


		RF	LogReg	SVM	XGB
European	Accuracy	0.76	0.72	0.66	0.72
	F1 Score	0.79	0.73	0.67	0.75
	AUROC	0.76	0.74	0.67	0.73
	AUPRC	0.76	0.75	0.69	0.74
Chinese	Accuracy	0.67	0.72	0.69	0.61
	F1 Score	0.56	0.60	0.54	0.48
	AUROC	0.65	0.68	0.65	0.58
	AUPRC	0.49	0.53	0.50	044
Combined	Accuracy	0.66	0.62	0.66	0.63
	F1 Score	0.63	0.50	0.64	0.62
	AUROC	0.66	0.62	0.66	0.63
	AUPRC	0.60	0.58	0.60	0.58

4.The table shows the **metrics of the tuned RF, LR, XGB and SVM classifie**r There is **no significant difference** between models.

CONCLUSION

The most important features identified by all classifiers for the European and Chinese data sets. Most of them have been verified by previous literature and can be considered as biomarkers



European	Chinese
Alistipes inops	*Acidaminococcus sp_CAG_542*
Ruminococcaceae bacterium_D16	*Prevotella bivia*
Alistipes shahii	*Lactobacillus mucosae*
Faecalibacterium prausnitzii	
Roseburia sp_CAG_182	

For the combined data the genera ***Clostridiales, Clostridium, Lactobacillus and Roseburia*** are of importance

Limitations and Future Work :

1. Further investigation of genus level relative abundances may provide more reliable results.
2. Further hyperparameter tuning must be conducted to improve performance of classifiers
3. Although ML methods may not be accurate enough to predict T2D, biomarkers may assist clinicians to make diagnosis.

Reference:

- Qin, Junjie, et al. "A metagenome-wide association study of gut microbiota in type 2 diabetes." Nature 490.7418 (2012): 55-60
- Karlsson, Fredrik H., et al. "Gut metagenome in European women with normal, impaired and diabetic glucose control." Nature 498.7452 (2013): 99-103.