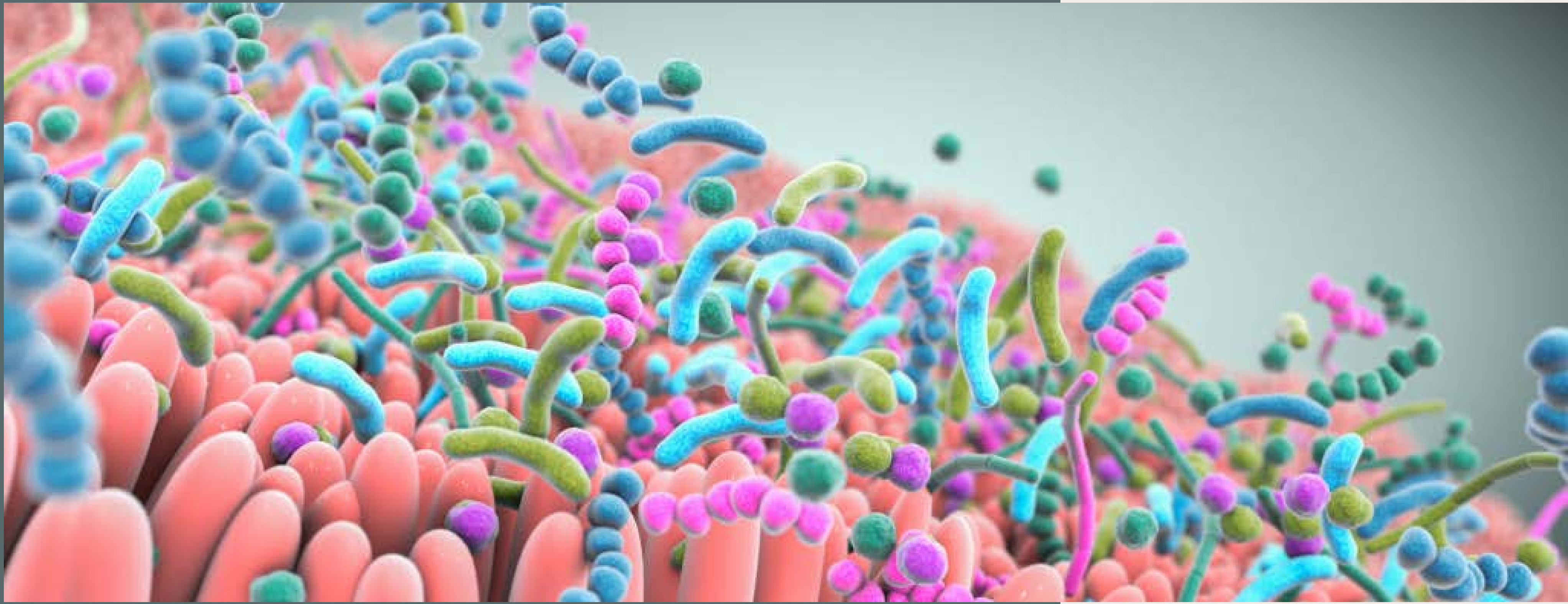


Ph.D. Miodrag Cekikj

"Ss. Cyril and Methodius" University in Skopje
FACULTY OF COMPUTER
SCIENCE AND ENGINEERING

Application of Machine Learning algorithms in modeling and understanding the role of the Microbiome in the Colorectal Cancer diagnosis and therapy

CA 18131 ML4Microbiome

APC Microbiome Ireland & ML4Microbiome Conference

University College Cork (UCC), Ireland

08/06/2023

Acknowledgements

COST Action CA18131 - Statistical and machine learning techniques in human microbiome studies, supported by the **European Cooperation in Science and Technology**

Faculty of Computer Science & Engineering, Ss. Cyril and Methodius University in Skopje, Macedonia

Biostatistics and Medical Informatics Department, **Acibadem Mehmet Ali Aydinlar University**, Istanbul, Turkey

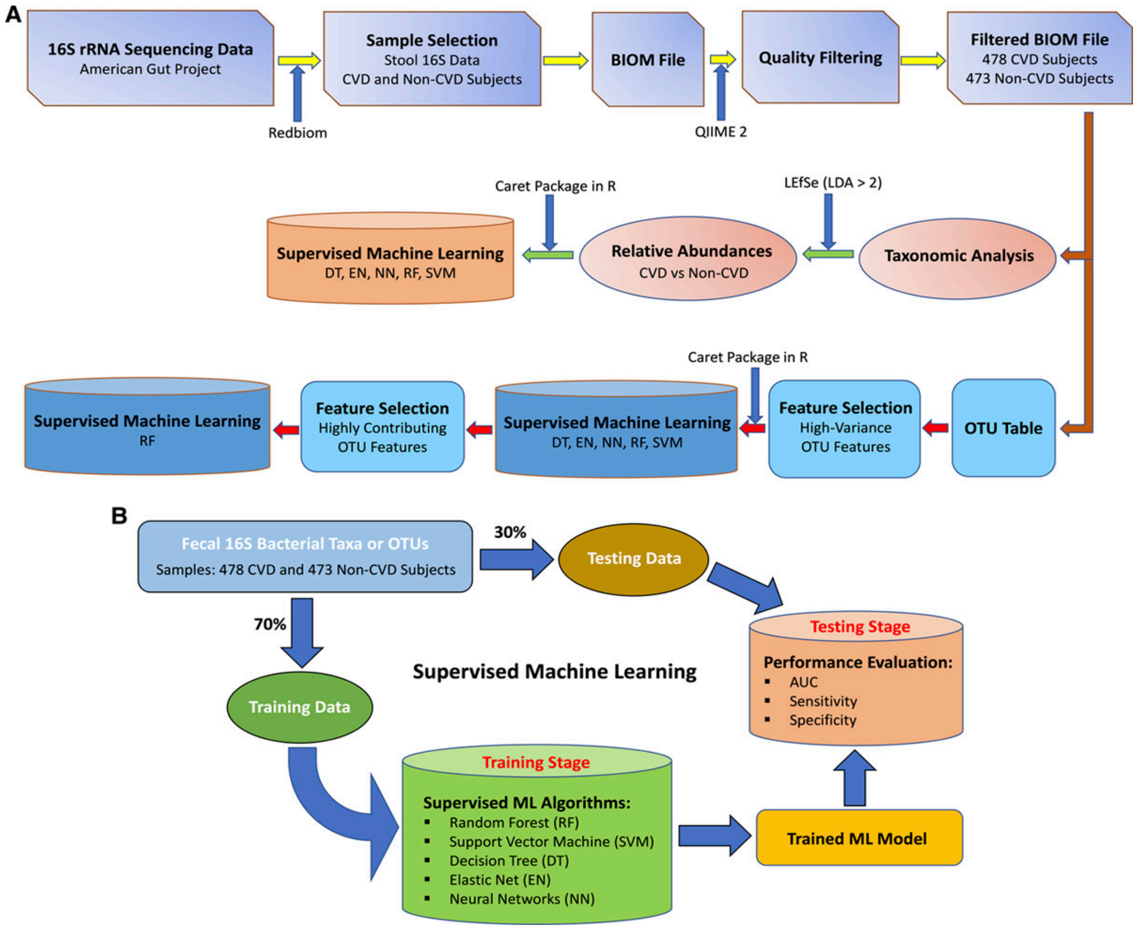
MSCA ITN Cell2Cell fellowship to L.L.E. project

MDPI Applied Sciences Journal
(the Section Applied Biosciences and Bioengineering)

Appl. Sci. 2022, 12(9), 4094; https://doi.org/10.3390/app12094094

@Sezerman Lab

‘De facto’ research methodology



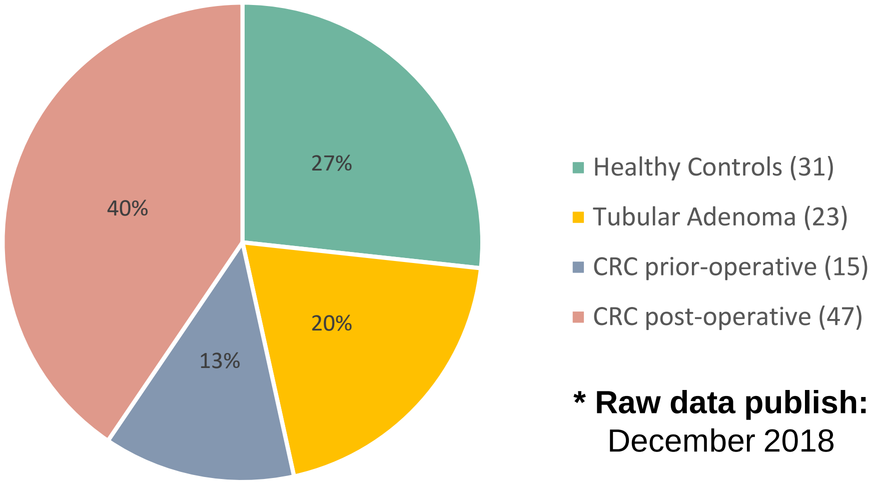
A. Overall lab & bioinformatics analysis

B. Machine Learning techniques

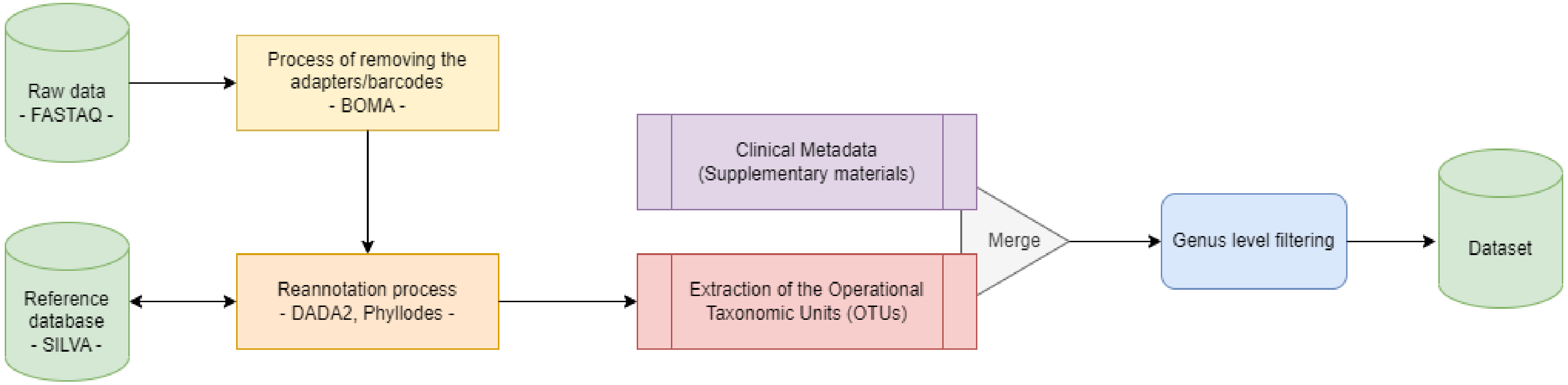
https://www.ahajournals.org/doi/full/10.1161/HYPERTENSIONAHA.120.15885

Dataset

- **116 individual microbiome samples** wrapped within **3603 Amplicon Sequence Variant (ASVs) units** phylogenetically defined in **<u>259 unique genera</u>.**
- Avoiding **data`s taxonomical bias** with <u>reannotation</u> of the raw reads against **updated bacterial references**.



*** Raw data publish:** December 2018

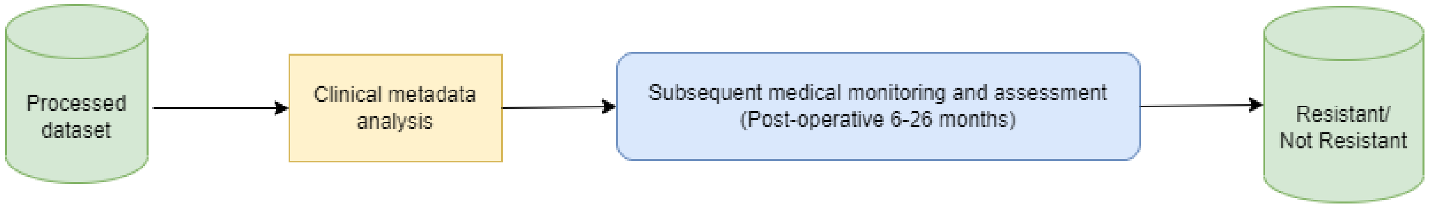


SILVA 138.1–16s reference db (latest reference database update on 27 August 2020).

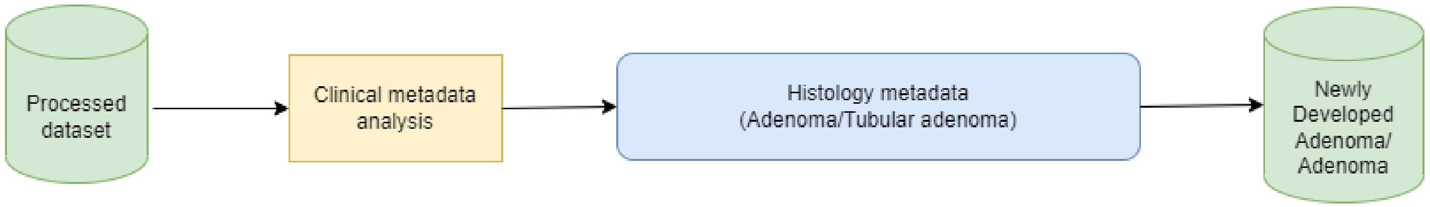
* Y. Jin et al., “Gut microbiota in patients after surgical treatment for colorectal cancer”, Environment Microbiology, vol. 21, no. 2, pp. 772–783, Feb. 2019, doi: 10.1111/1462-2920.14498.

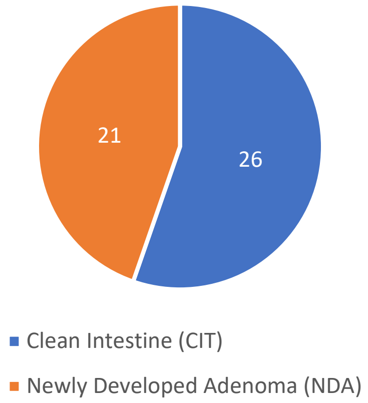
Case studies data

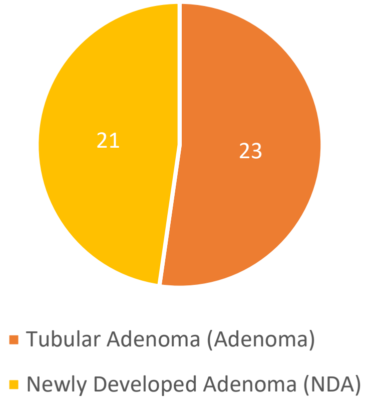
I. **Drug resistance mechanism** (immunotherapy effect)


Processed dataset
Clinical metadata analysis
Subsequent medical monitoring and assessment (Post-operative 6-26 months)
Resistant/ Not Resistant

II. **CRC Carcinogenesis** (histology-based study)

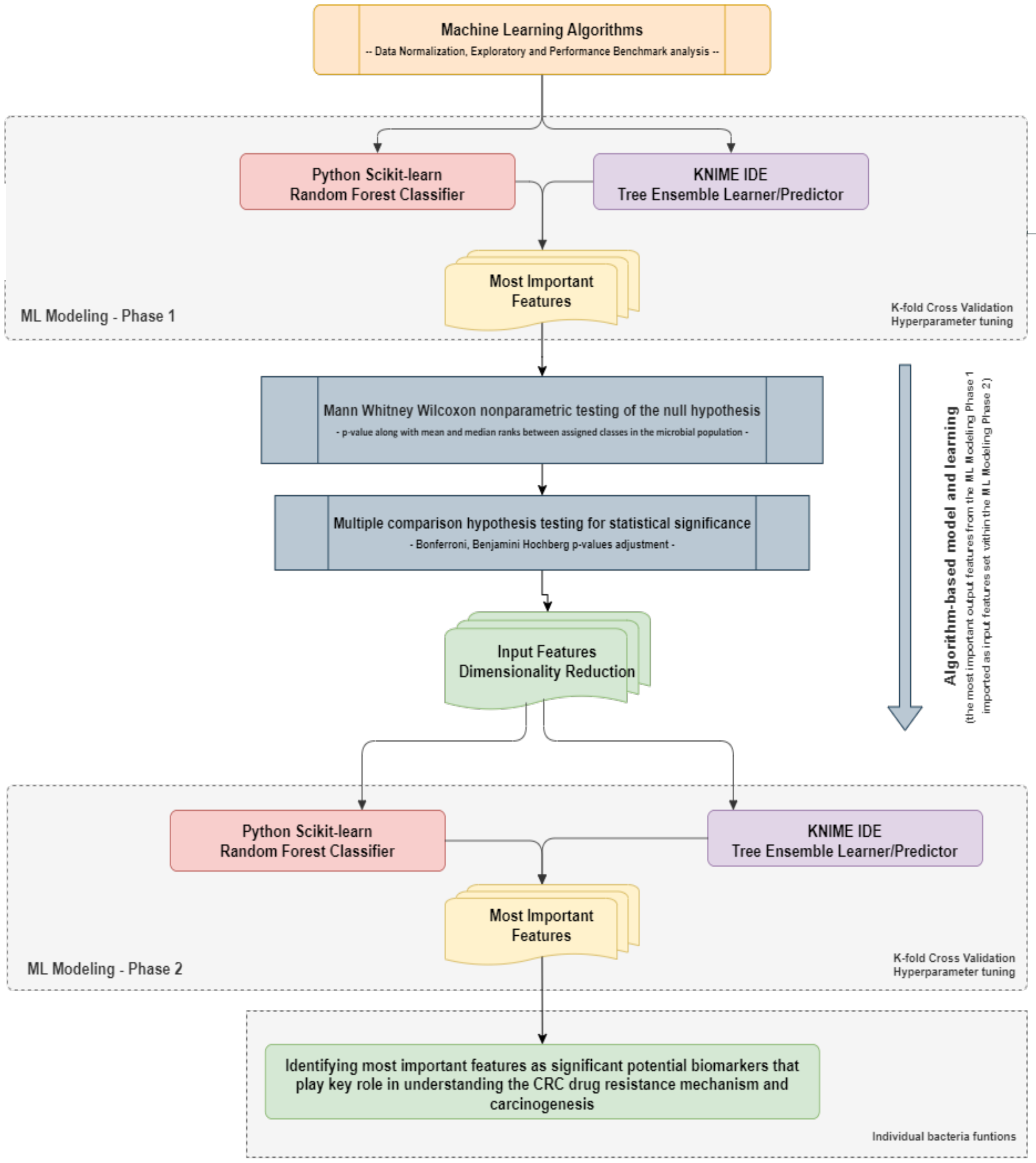

Processed dataset
Clinical metadata analysis
Histology metadata (Adenoma/Tubular adenoma)
Newly Developed Adenoma/ Adenoma





Bioinformatics Methodology

- Machine learning and statistics as a **supervised learning approach** to examine the biological features.
- Reduce **and semantically interpret the input set** by designing the modeling process into two **subsequent stages**.
- Algorithm hyperparameter tuning for **n_estimators**, **max_depth**, and **max_feature** (RandomizedSearchCV/GridSearchCV).
- **Analytical feature reduction and engineering** over, for example, the recursive features elimination (RFE) procedure.
- **Statistical and non-parametric data testing** and analysis to examine the abundance within the different classes and find more data insights for further biological evaluations and findings.



Machine Learning Algorithms

-- Data Normalization, Exploratory and Performance Benchmark analysis --

Python Scikit-learn
Random Forest Classifier

KNIME IDE
Tree Ensemble Learner/Predictor

Most Important Features

ML Modeling - Phase 1

K-fold Cross Validation
Hyperparameter tuning

Mann Whitney Wilcoxon nonparametric testing of the null hypothesis

- p-value along with mean and median ranks between assigned classes in the microbial population -

Multiple comparison hypothesis testing for statistical significance

- Bonferroni, Benjamini Hochberg p-values adjustment -

Input Features
Dimensionality Reduction

Algorithm-based model and learning

(the most important output features from the ML Modeling Phase 1
imported as input features set within the ML Modeling Phase 2)

Python Scikit-learn
Random Forest Classifier

KNIME IDE
Tree Ensemble Learner/Predictor

Most Important Features

ML Modeling - Phase 2

K-fold Cross Validation
Hyperparameter tuning

Identifying most important features as significant potential biomarkers that play key role in understanding the CRC drug resistance mechanism and carcinogenesis

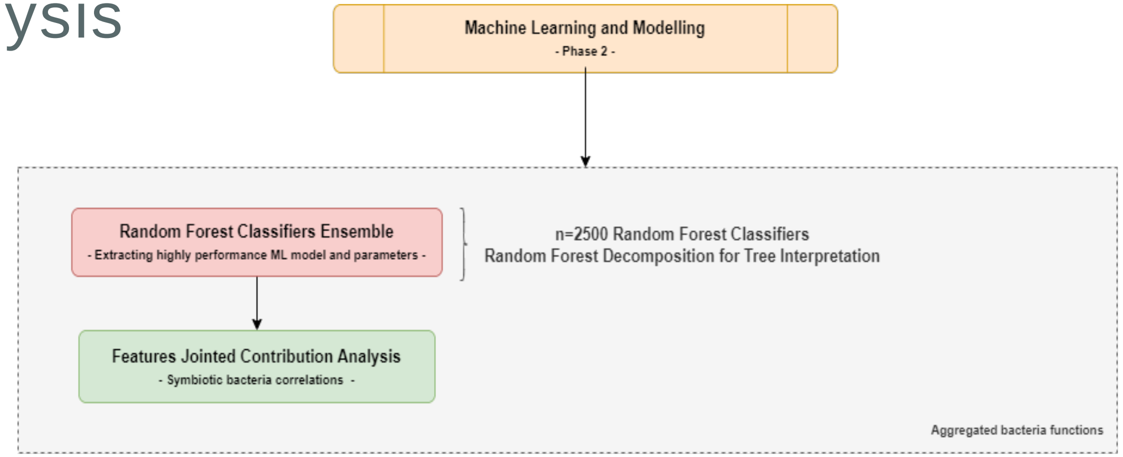
Individual bacteria funtions

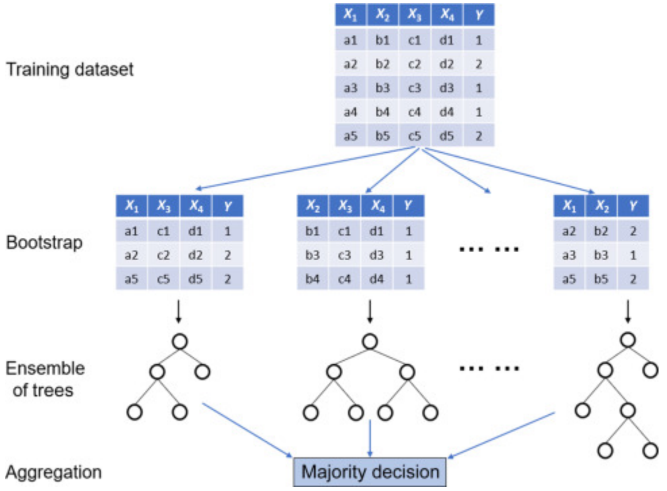
Aggregated features contribution analysis

- **Joint feature combinations**, providing a **combined overview of the model's predictability** corresponding to the resistance class.

- The aggregated contributions are lower than the individual ones but uncover additional data insights regarding the constitution of the entire trajectory along the algorithm's prediction path.

- **tree interpreter library (v.0.2.3)** - decomposing the prediction contribution for the individual predictions and aggregated them for the whole data set (using the **aggregated contributions convenience method**).


Machine Learning and Modelling
- Phase 2 -
Random Forest Classifiers Ensemble
- Extracting highly performance ML model and parameters -
n=2500 Random Forest Classifiers
Random Forest Decomposition for Tree Interpretation
Features Jointed Contribution Analysis
- Symbiotic bacteria correlations -
Aggregated bacteria functions



X_1	X_2	X_3	X_4	Y
a1	b1	c1	d1	1
a2	b2	c2	d2	2
a3	b3	c3	d3	1
a4	b4	c4	d4	1
a5	b5	c5	d5	2

X_1	X_3	X_4	Y
a1	c1	d1	1
a2	c2	d2	2
a5	c5	d5	2

X_2	X_3	X_4	Y
b1	c1	d1	1
b3	c3	d3	1
b4	c4	d4	1

X_1	X_2	Y
a2	b2	2
a3	b3	1
a5	b5	2

Results

ML Modelling - Screening Phase

ML Algorithm	Overall Accuracy *
Naïve Bayes	0.429
Logistic Regression	0.425
K-Nearest Neighbors	0.325
Support Vector Machine	0.497
Decision Tree	**0.764**

* The overall algorithm accuracy was selected as the main algorithm selection indicator.

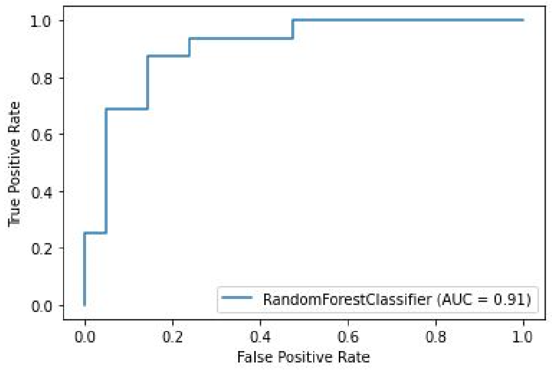
- **Drawbacks: Naïve Bayes** (all features are independent), **Logistic regression** (linearity between the dependent variable and the independent variables), **KNN** (high dimensionality & the sensitivity of choosing the neighbors based on the distance criteria).
- **Decision Tree** with **‘gini’ attribute selection measure** in correlation with the **‘best’ splitter** as splitting strategy approach.
- Additional benefits: **DT comprehensibility** & **taking advantage of the tree-related majority voting (Random Forest)**

CRC Drug-resistance Mechanism Results
ML Modelling - Main Phase

General ML modeling performance metrics for the resistant and non-resistant CRC post-operative individuals' group

Environment	ML Algorithms	Normalization/Scaling	Accuracy	Sensitivity	Specificity
Python Scikit-learn	RFC (P1)	Standard Scaler	0.9	1.000	0.833
Python Scikit-learn	RFC (P1)	Z-Score Normalizer	0.9	1.0	0.75
KNIME	TEL (P1)	Z-Score Normalizer	0.833	0.778	1.0
Python Scikit-learn	**RFC (P2)**	**Standard Scaler**	**0.917**	**1.000**	**0.833**
KNIME	TEL (P2)	Z-Score Normalizer	0.9	1.000	0.8

Aggregated measure of performance of a binary classifier on all possible **threshold values**



* RFC - Scikit-learn random forest classifier, TEL - Tree ensemble learner, P1 - Phase 1 ML modeling, P2 - Phase 2 ML modeling.

** **Sensitivity** = **Recall** = TP/(TP+FN) - correctly predicted by the model; **Specificity** = True Negative Rate = TN/(TN+FP)

- **First phase:** **n_estimators = 55**, **max_depth = 5, max_features = 3**, **cross-validation value of 25% test data** using the stratified sampling by additionally introduced 'resistance' target feature.
- **Second phase:** **n_estimators = 25**, **max_depth = 4**, **max_features = 3**, cross-validation value of 25% test data, Area under the curve (AUC) = 0.91 (reasonable discriminated ability to classify).

CRC Drug-resistance Mechanism Results
Aggregated Features Contribution Analysis

- ***Enterococcus***, ***Blautia***, ***Subdoligranulum***, and ***Escherichia-Shigella*** were mostly observed contributing to the resistant group.

- ***Enterococcus*** is identified in correlation to ***Haemophilus***, ***Intestinibacter***, ***Ruminococcus***, ***Lachnoclostridium***, ***Weissella***, ***Coprococcus***, and ***Senegalimassilia***.

- ***Blautia*** is commonly significant with ***Paraprevotella***, ***Subdoligranulum***, ***Oxalobacter***, and ***TM7x*** genera.

- ***Escherichia-Shigella*** is mostly observed in aggregated relation to ***Subdoligranulum***, ***Coprococcus***, ***Gemella***, and ***Negativibacillus***.

Aggregated Bacteria	**‘Resistance’ Contribution**
[‘Escherichia-Shigella’, ‘Subdoligranulum’, ‘Gemella’, ‘Negativibacillus’]	0.00770053
[‘Blautia’, ‘TM7x’] [‘	0.0061875
[‘Escherichia-Shigella’, ‘Coprococcus’, ‘Lachnospiraceae UCG-010’, ‘Family XIII UCG-001’]	0.00555556
[‘Terrisporobacter’, ‘Weissella’, ‘Slackia’]	0.00538462
[‘Enterococcus’, ‘Haemophilus’, ‘UCG-005’]	0.005
[‘Intestinibacter’, ‘Enterococcus’, ‘Lachnospiraceae NC2004 group’, ‘Lachnoclostridium’]	0.0047138
[‘Coprococcus’, ‘Megasphaera’, ‘Parasutterella’, ‘UCG-002’]	0.0045
[‘Streptococcus’, ‘Phascolarctobacterium’, ‘Paraprevotella’, ‘Dubosiella’]	0.00403846
[‘Subdoligranulum’, ‘Blautia’, ‘Paraprevotella’, ‘Oxalobacter’]	0.00317853
[‘Subdoligranulum’, ‘Butyrivibrio’]	0.00307692
[‘Lachnospiraceae UCG-010’, ‘Barnesiella’]	0.00235897
[‘Blautia’, ‘Oxalobacter’] [‘	0.00231884
[‘Clostridium sensu stricto 1’, ‘Flavonifractor’, ‘Agathobacter’, ‘Butyricimonas’]	0.00227193
[‘Flavonifractor’, ‘Agathobacter’, ‘Butyricimonas’, ‘Anaerofustis’]	0.00222222
[‘[Eubacterium] ruminantium group’, ‘[Eubacterium] eligens group’, ‘Moryella’]	0.00198413

Aggregated bacteria significance contributions to the **resistant class**

CRC Drug-resistance Mechanism Results
Biological analysis and interpretation

- The **enterotoxigenic *Bacteroides* bacteria** has a **critical impact on the CRC development and proliferation** considering their biofilm production for colonization that results in a **series of inflammatory reactions** that encourages chronic intestinal inflammation and tissue damage.

- The ***Alistipes* bacteria** is **living in symbiosis with the *Bacteroides* species because both are resistant to vancomycin, kanamycin, and colistin**. These two species have similar pathways for amino acid fermentation supporting colon inflammation and adenoma development.

- The ***Barnesiella*** species shows high correlation with the **non-resistant group; but** its metabolites indicate infiltration of interferon-γ-producing γδT cells in cancer tissues.

Genus	Research findings	p-value
Barnesiella	↑	0.0069
Alistipes	↑	0.0017
Intestinibacter	↑	0.038
Flavonifractor	↓	0.04
Akkermansia	↑	0.041
[Ruminococcus] torques group	↓	0.043
Streptococcus	↓	0.021
Butyricimonas	↑	0.022
Eggerthella	↓	0.024
Escherichia-Shigella	↓	0.026
Anaerovoracaceae	↑	0.027
Negativibacillus	↑	0.031
Leuconostoc	↓	0.034
Ruminococcus	↓	0.0017
Oscillospiraceae	↑	0.0034
Bacteroides	↓	0.0087
Clostridium sensu stricto 1	↑	0.015

↑ Increased presence and impact in non-resistant samples

↓ Reduced presence and impact in non-resistant samples

Further Scientific Actions

- The established methodology can also be used for **unseen microbiome data that can help oncologists decide on treatment and post-treatment strategy for immunotherapy and drug resistance understandings**.
- Improve the **symbiotic bacterial analysis for providing a combined overview of the model's predictiveness and uncovering additional data correlations**.
- General **microbiome-agnostic model** (reinforcement learning approach).
- **Blockchain** utilization in the picture.

ACIBADEM ÜNIVERSITESI
PARAFILM
ACIBADEM ÜNIVERSITESI

Thank you for your attention!

> „The philosophers have only *interpreted* the world, in various ways. The point, however, is to *change* it.
>
> — Karl Marx, Eleven Theses on Feuerbach