VOC analysis

Marcos Fabietti

mifabietti@outlook.com

14/09/2024

Introduction

In this repository we explore 3 datasets of VOCs in order to showcase analytical methods that can be used to gain insight into their use. These include:

- **Breath Biopsy® OMNI® – Example Dataset:** This dataset provides an example of the data generated from samples of human breath VOCs by Breath Biopsy OMNI. It illustrates the kind of feature table that is provided as a basic output from Breath Biopsy studies.
- **A Clinical Breathomics Dataset:** This study entailed a comprehensive GC–MS analysis conducted on 121 patient samples to generate a clinical breathomics dataset. This dataset cataloged volatile organic compounds (VOCs) from the breath of individuals with asthma, bronchiectasis, and chronic obstructive pulmonary disease.
- **Breast cancer-related VOCs:** This dataset was collected to evaluate the diagnostic performance of breath-omics to differentiate between benign and malignant breast lesions and assess the diagnostic performance of a multi-omics approach combining breath-omics, ultrasound radiomics, and clinic-omics via a nested cohort study.

Overview

Problem Statement

- Breath Biopsy® OMNI®: What insights can we gain from breath samples compared to "blank" samples?
- A Clinical Breathomics Dataset: Can we use breathomics to identify which respiratory disease a patient has, and which are the key VOCs?
- Breast cancer-related VOCs: Can we use breathomics to identify patients with malignant breast cancer compared to benign, and the key VOCs?

Background

- Breathomics is a field of study that analyzes exhaled breath to detect volatile organic compounds (VOCs) and their role in health. VOCs are chemical markers that reflect metabolic changes in the body, and analysing them can provide early disease detection. With non-invasive diagnostics, we can deliver a non-invasive, painless, and rapid diagnostic method that can identify disease signatures much earlier than conventional tests, which often require blood tests, biopsies, or imaging techniques.

Goal

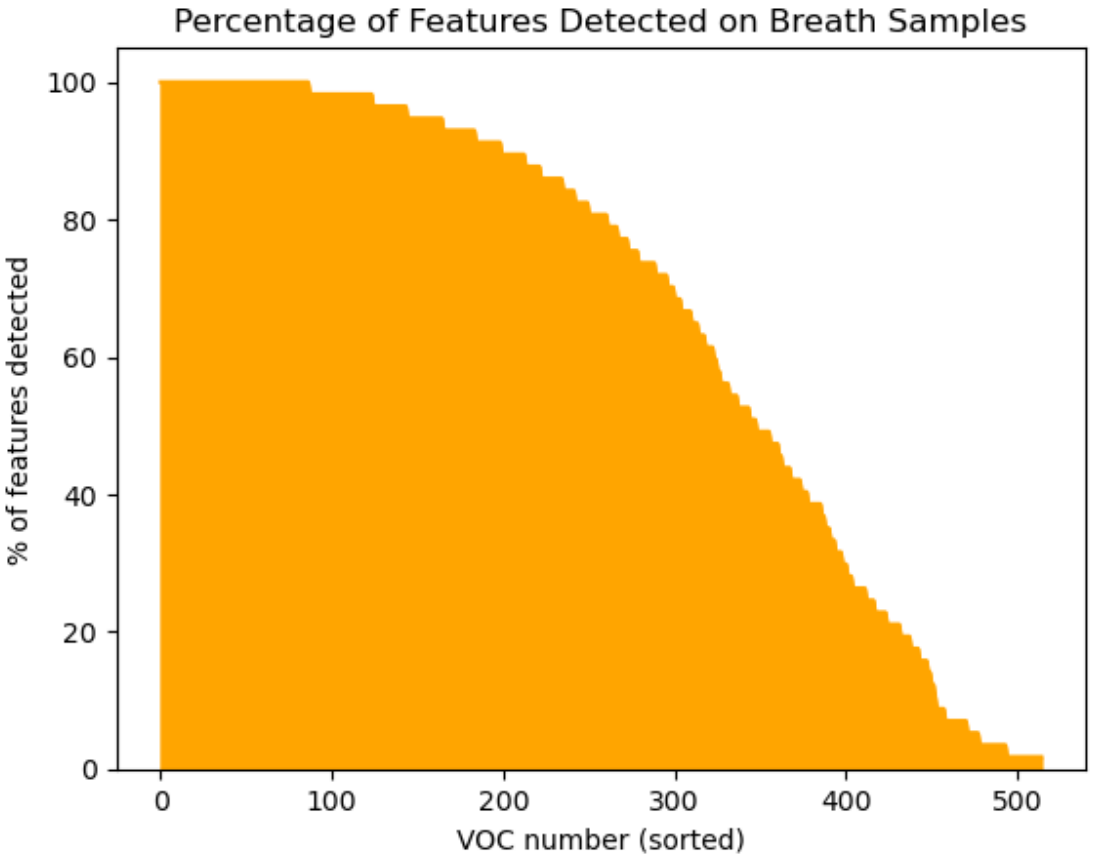
- Using advanced data modelling techniques, we want to make gain insights of VOCs and their diagnostics capabilities. Ultimately, we want to find ways we can use these predictions to help diagnose in a non-invasive manner.

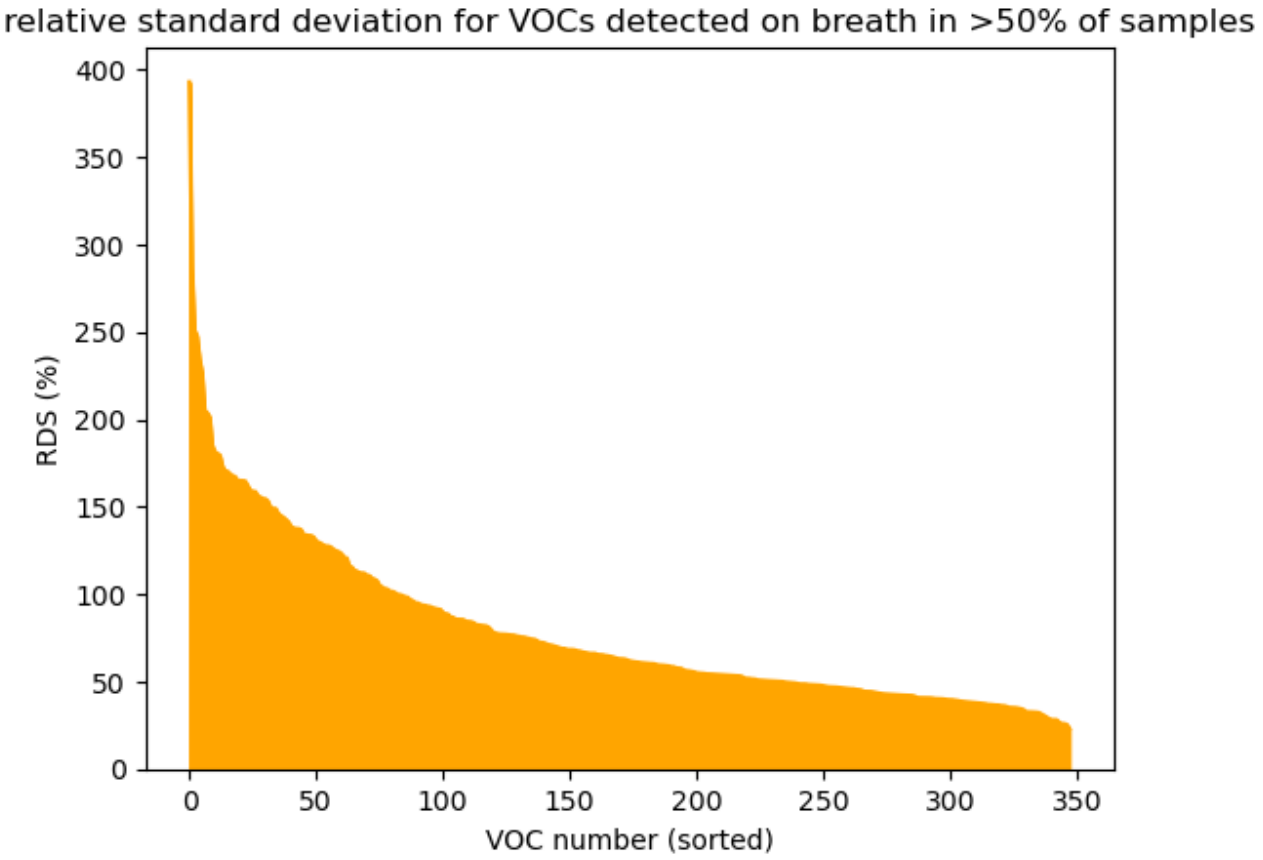
Methods

The datasets are fairly small, of 166 (Breath Biopsy® OMNI® – Example), 121 (A Clinical Breathomics Dataset) and 476 (Breast cancer-related VOCs) examples. For the second dataset we have "Disease" as a response variable, making it a multi-class problem of asthma/bronchiectasis/COPD, whereas the third dataset the response variable is "Label", a binary category of benign/malignant. While we could classify the examples of the first dataset between "breath" and "blank", the reference article indicated that a simple threshold of 3 standard deviations from the blank samples would suffice, thus we employed that. The second and third dataset have patient data in addition to the VOCs, and we kept them in order to evaluate if the former are stronger predictors than the latter.

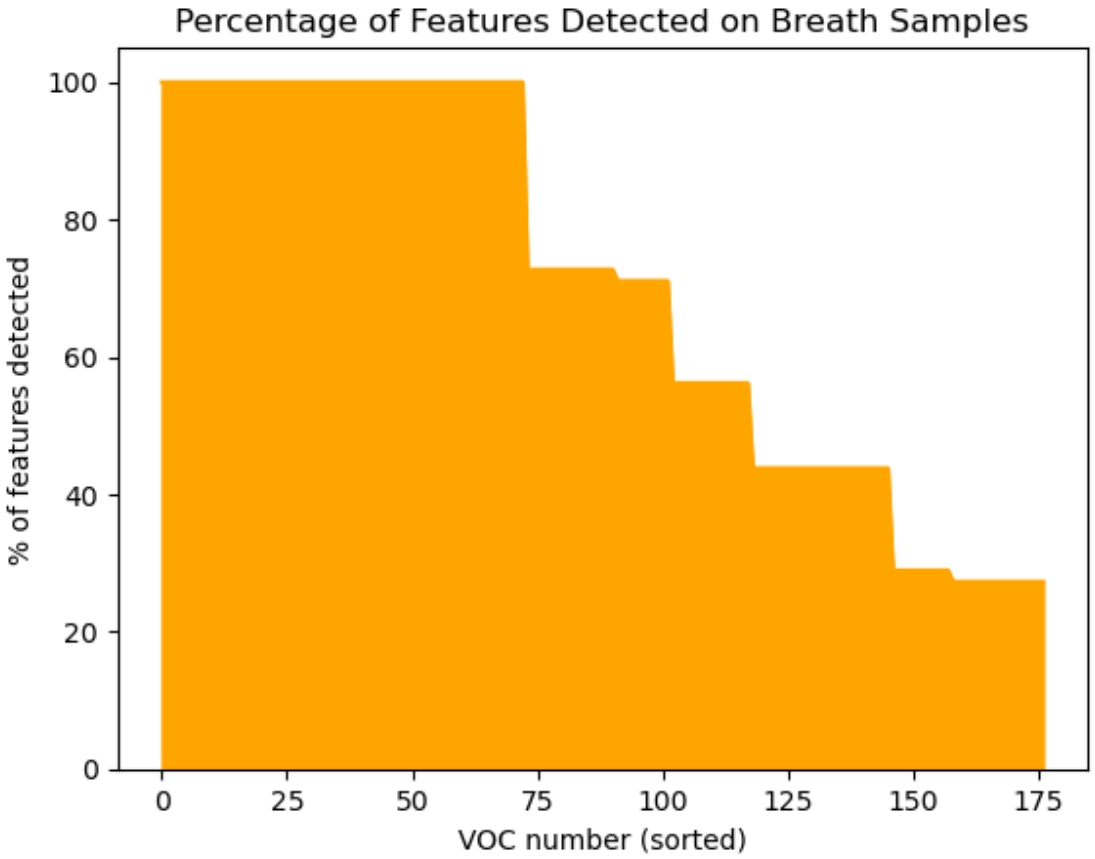
We replaced the missing data with the mean of each class where needed, RandomOversampled rows to match the number of examples to the majority class and scale each variable via standardisation. Subsequently, we did a 5-Kfold split in order to train the ML models including: Decision Tree-based, Random Forest, Gradient Boosting, K-Nearest Neighbours and Support Vector Machines. We judged our model based on Accuracy, Precision, Recall and F1 score in the binary classification, and just on accuracy for the multi-class classification. Lastly, we employed a Mann–Whitney U test to evaluate if there is a statistical difference between the two groups for each of the top 5 features highlighted by the SHAP values.

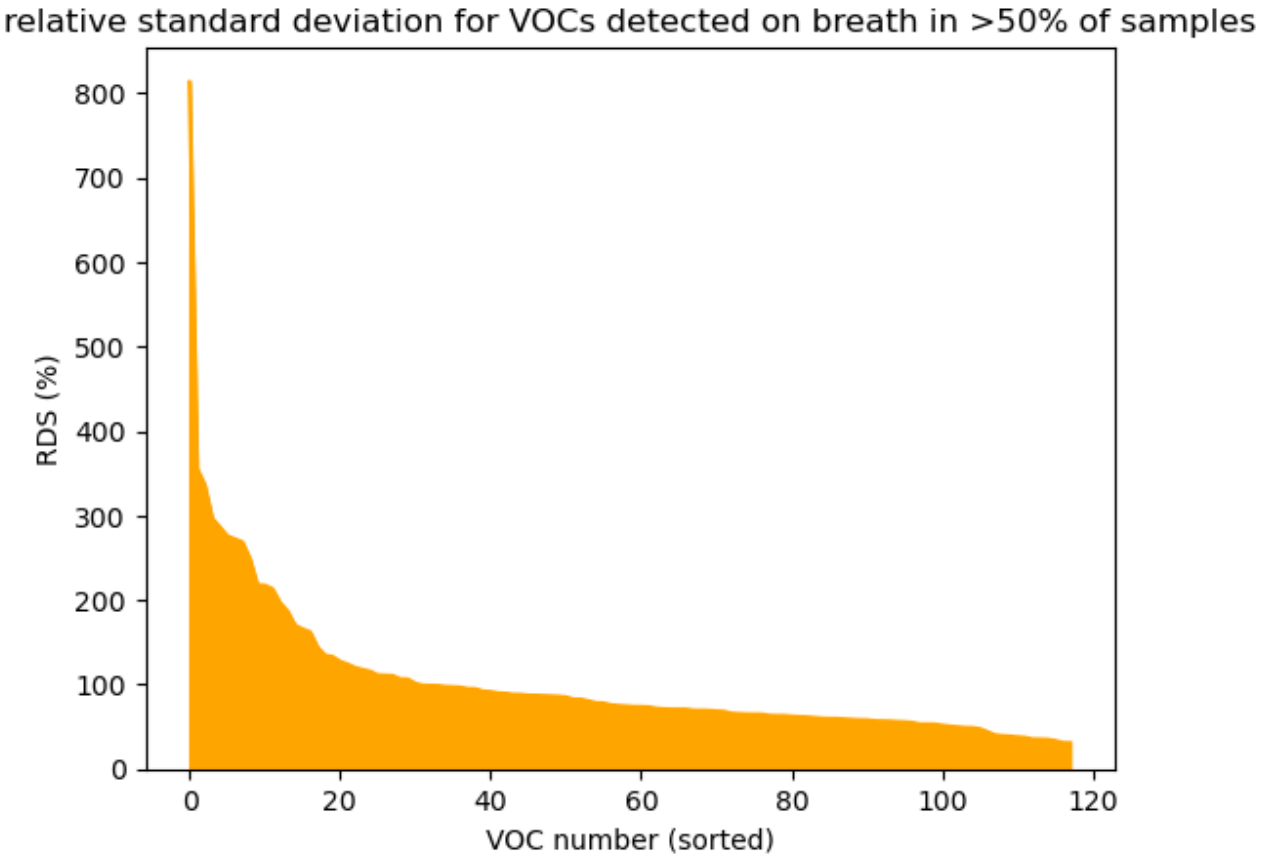
Breath Biopsy® OMNI® – Example Dataset



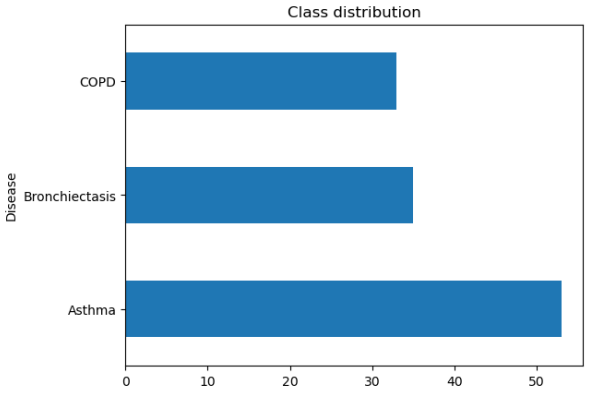


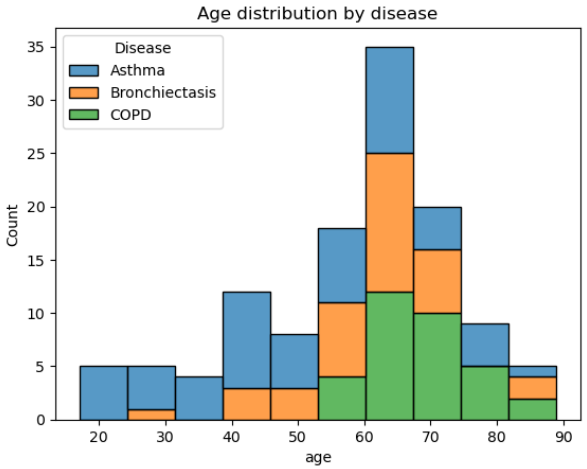
A Clinical Breathomics Dataset

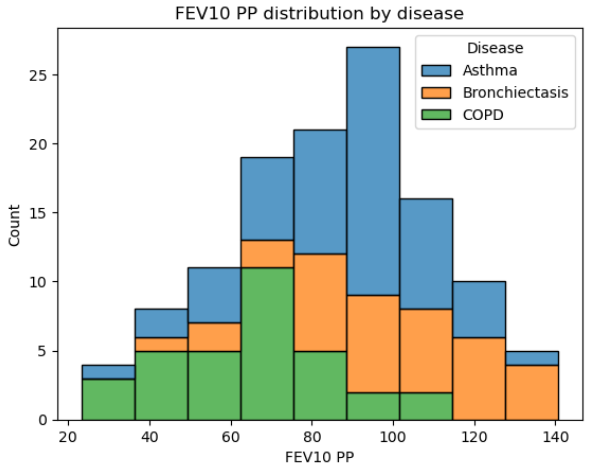


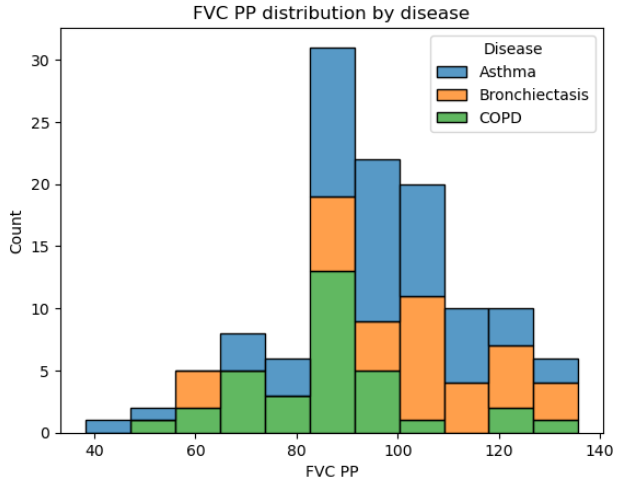


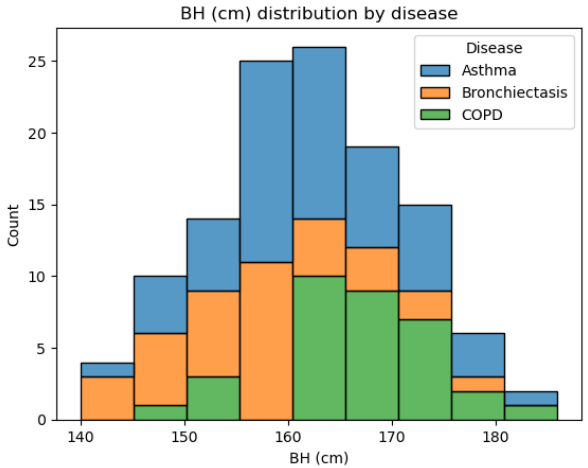
EDA

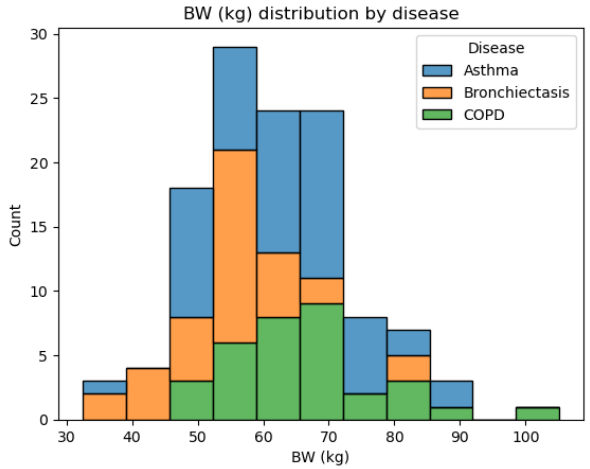

Class distribution
COPD
Bronchiectasis
Asthma
Disease
0
10
20
30
40
50

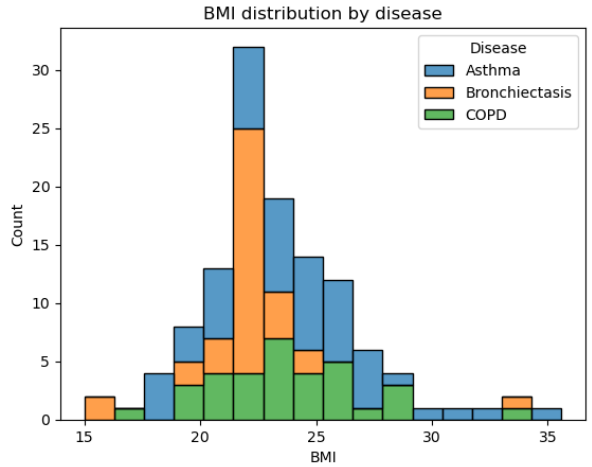

Age distribution by disease
Disease
Asthma
Bronchiectasis
COPD
Count
age

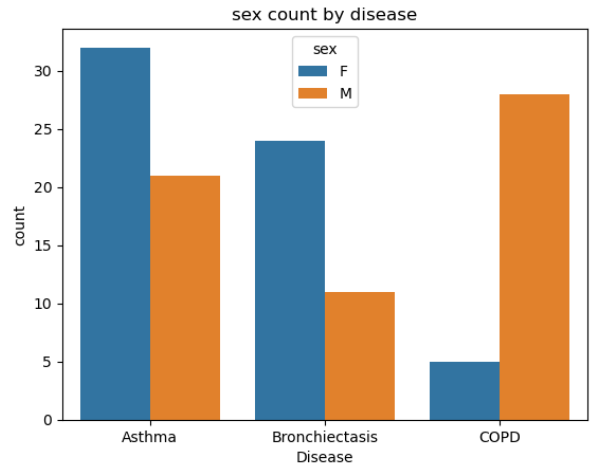

FEV10 PP distribution by disease
Disease
Asthma
Bronchiectasis
COPD
Count
FEV10 PP


FVC PP distribution by disease
Disease
Asthma
Bronchiectasis
COPD
Count
FVC PP

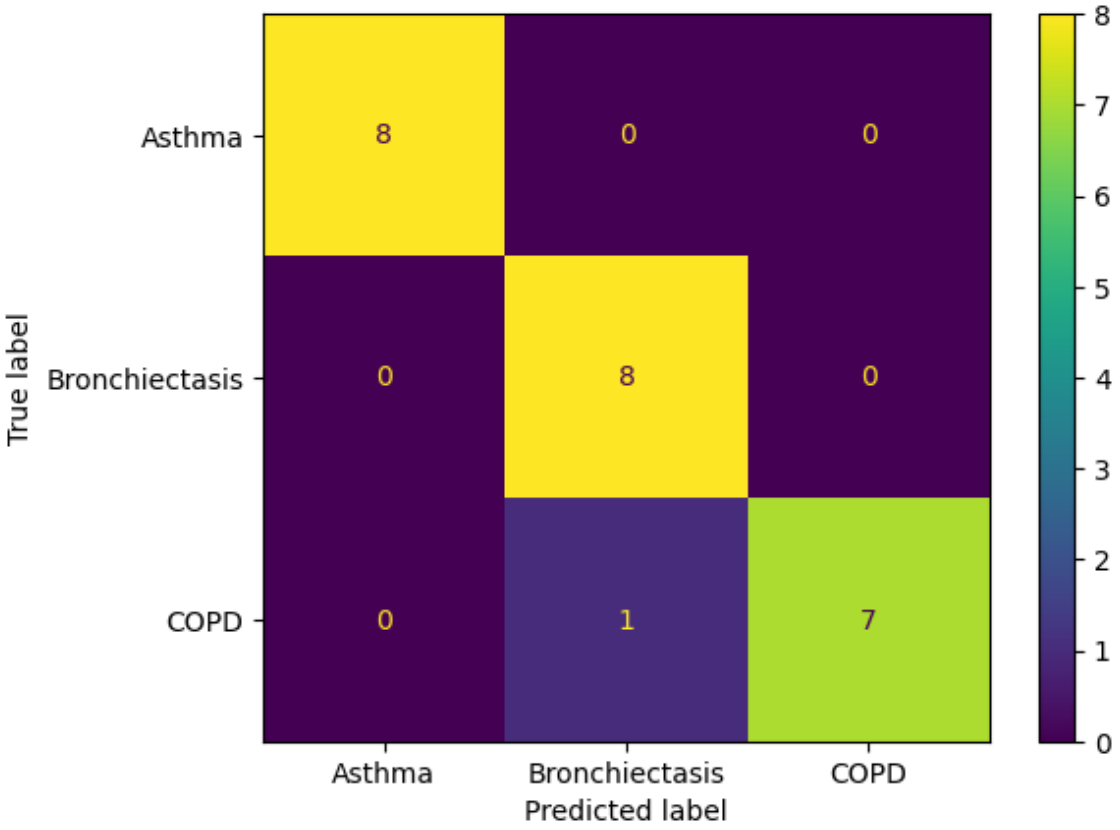

BH (cm) distribution by disease
Disease
Asthma
Bronchiectasis
COPD
Count
BH (cm)


BW (kg) distribution by disease
Disease
Asthma
Bronchiectasis
COPD
Count
BW (kg)


BMI distribution by disease
Disease
Asthma
Bronchiectasis
COPD
Count
BMI

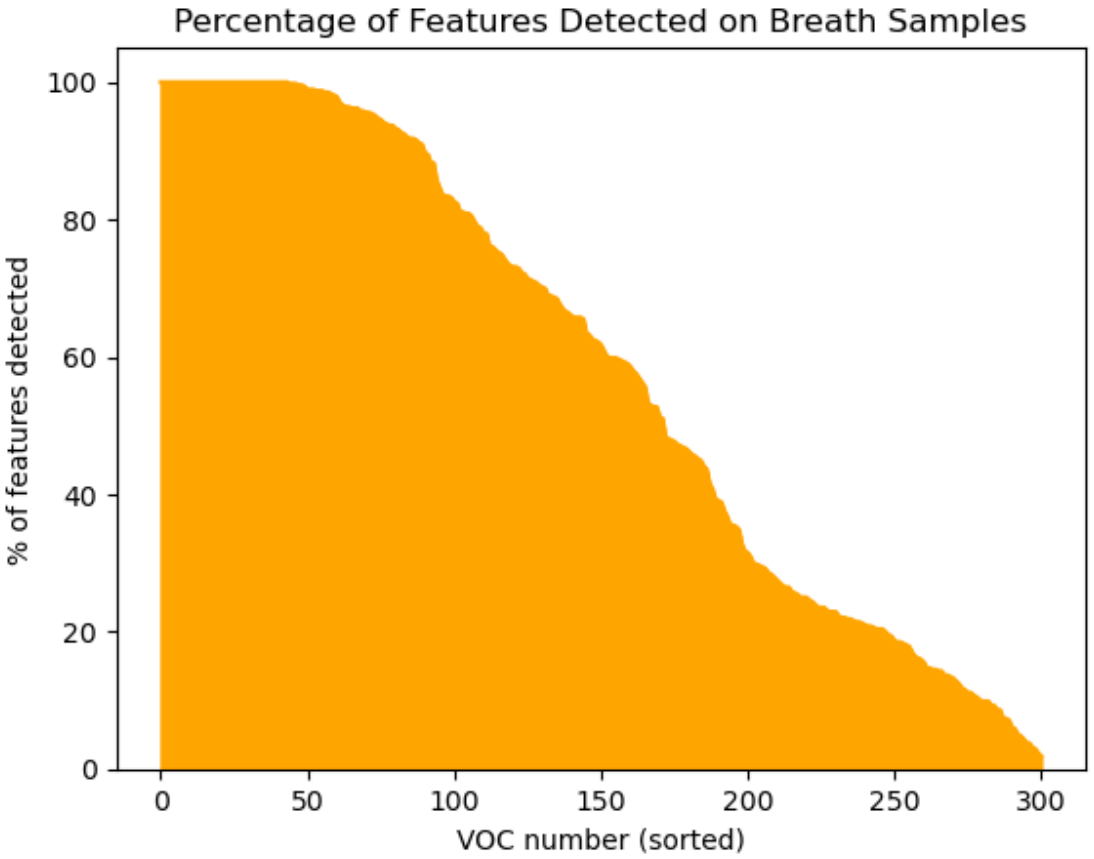

sex count by disease
sex
F
M
count
Asthma
Bronchiectasis
COPD
Disease

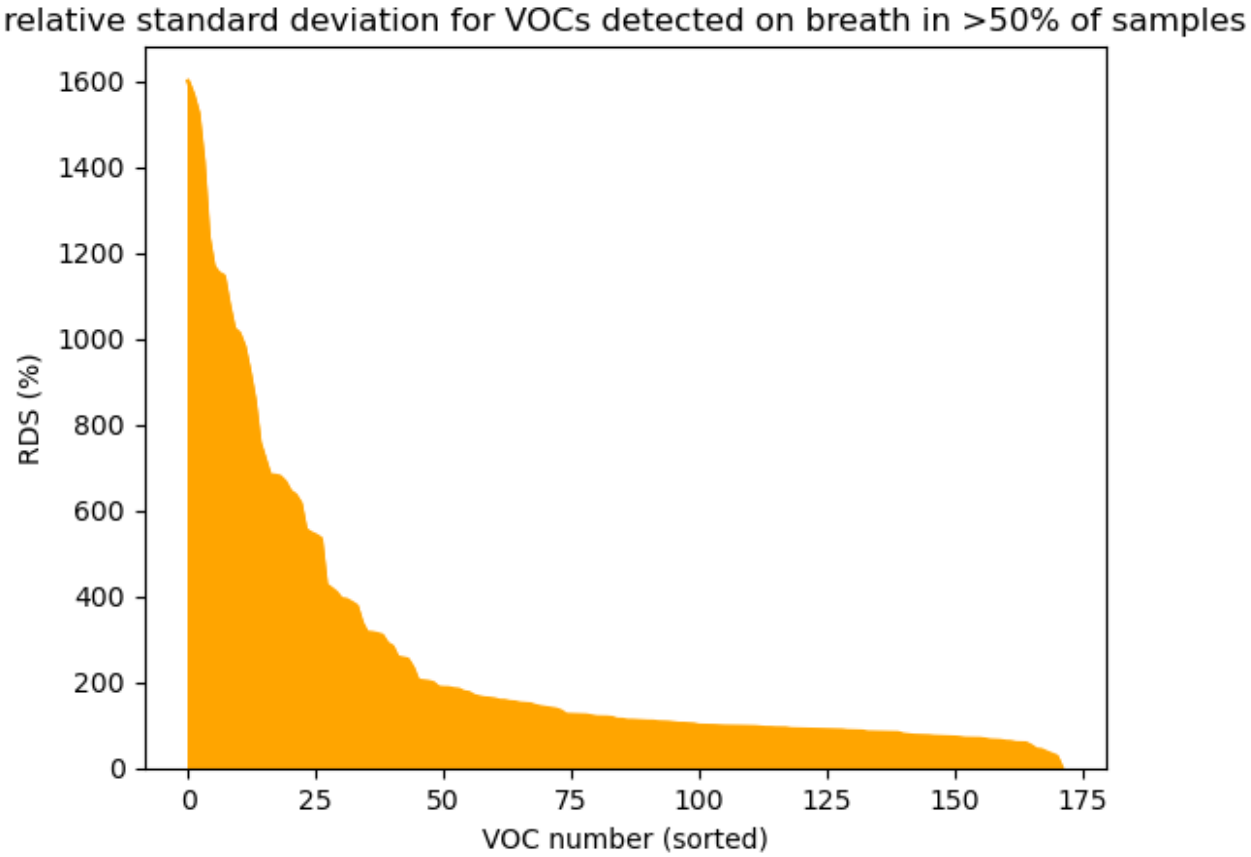
Results



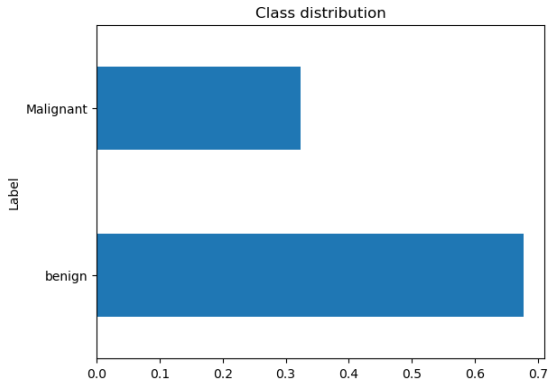
m/z	Potential VOCs	CAS number	Molecular weight	Molecular formula	feature importance	Biomarker Class	Max p-value across test against other classes
138	2-pentylfuran	3777-69-3	138.2069	C9H14O	0.098637	COPD	1.059131e-05
519	2,2,4,4,6,6,8,8,10,10,12,12,14,14-tetradecamethyl-1,3,5,7,9,11,13-heptaoxa-2,4,6,8,10,12,14-heptasilacyclotetradecane	107-50-6	519.08	C14H42O7Si7	0.064379	Asthma	8.390869e-15
226	hexadecane	544-76-3	226.445	C16H34	0.038067	Asthma	8.390869e-15
120	1-ethyl-4-methylbenzene	622-96-8	120.1916	C9H12	0.033673	Asthma	1.666728e-05
128	azulene	275-51-4	128.1705	C10H8	0.021945	Bronchiectasis	6.136696e-11

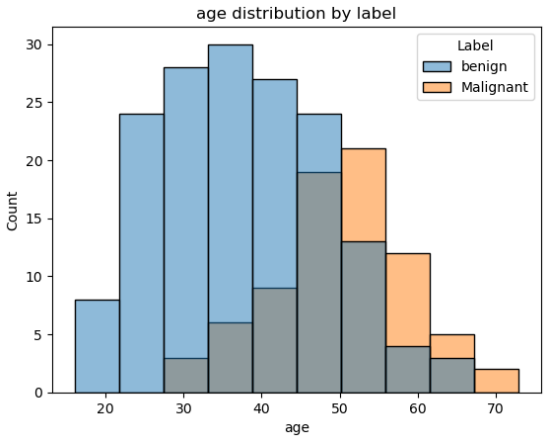
Breast cancer-related VOCs

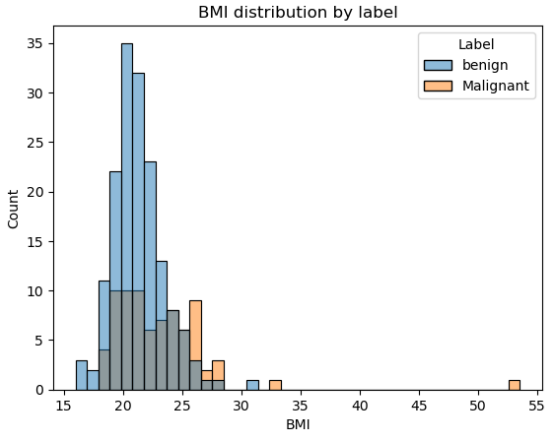


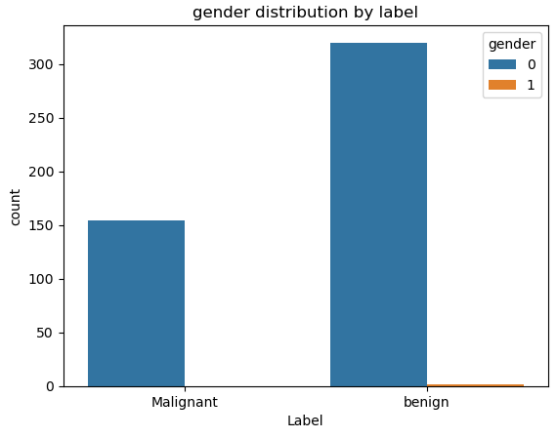


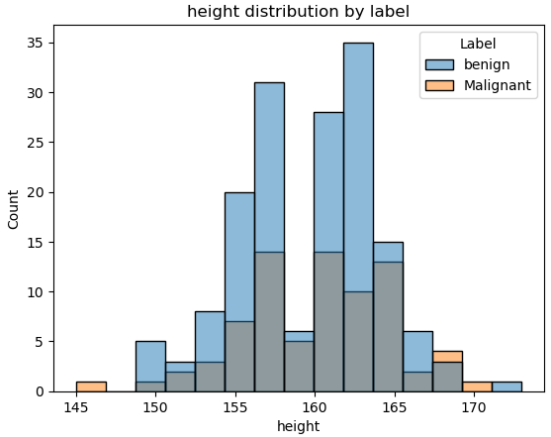
EDA

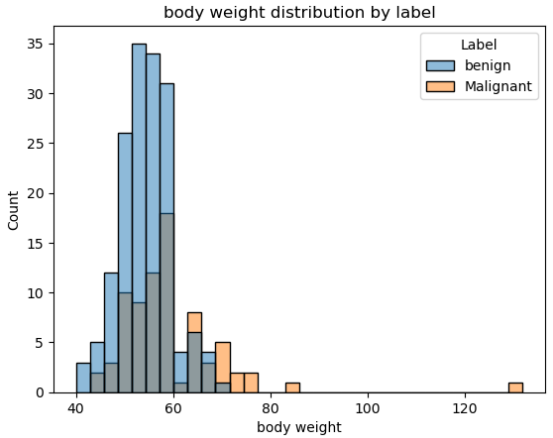

Class distribution
Malignant
benign
Label
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7


age distribution by label
Label
benign
Malignant
Count
age
20
30
40
50
60
70

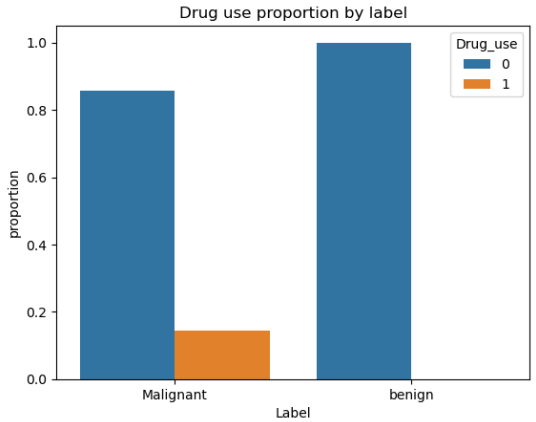

BMI distribution by label
Label
benign
Malignant
Count
BMI
15
20
25
30
35
40
45
50
55

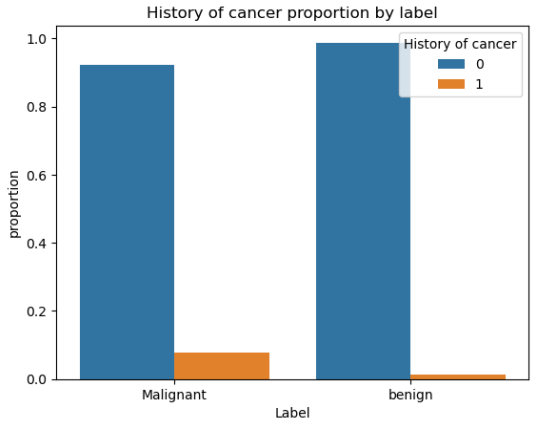

gender distribution by label
gender
0
1
count
Malignant
benign
Label
0
50
100
150
200
250
300

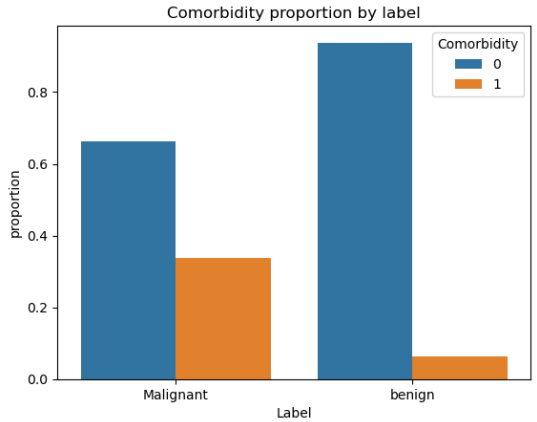

height distribution by label
Label
benign
Malignant
Count
height
145
150
155
160
165
170

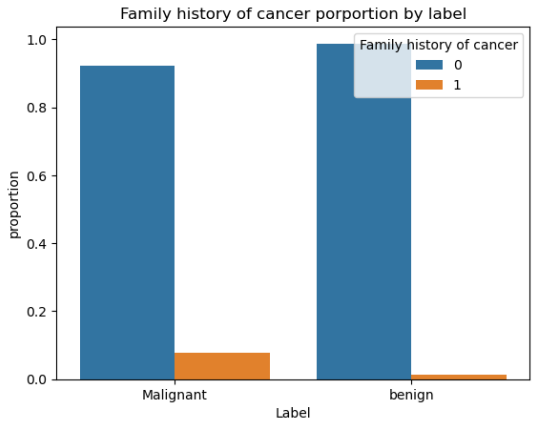

body weight distribution by label
Label
benign
Malignant
Count
body weight
40
60
80
100
120

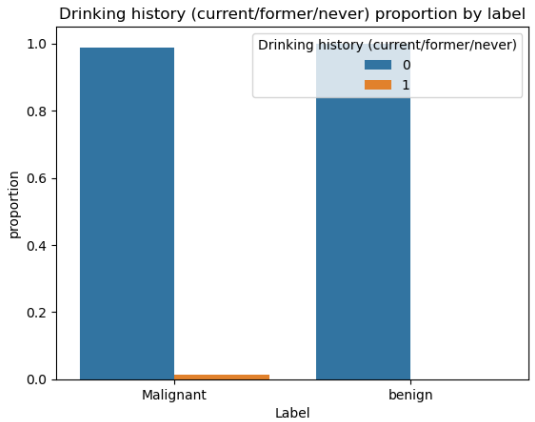
EDA

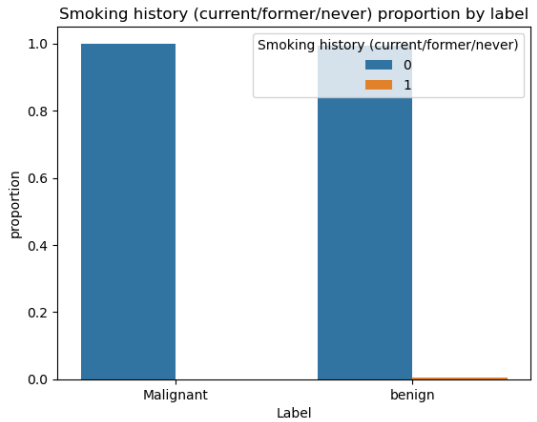

Drug use proportion by label
Drug_use
0
1
proportion
0.0
0.2
0.4
0.6
0.8
1.0
Malignant
benign
Label


History of cancer proportion by label
History of cancer
0
1
proportion
0.0
0.2
0.4
0.6
0.8
1.0
Malignant
benign
Label

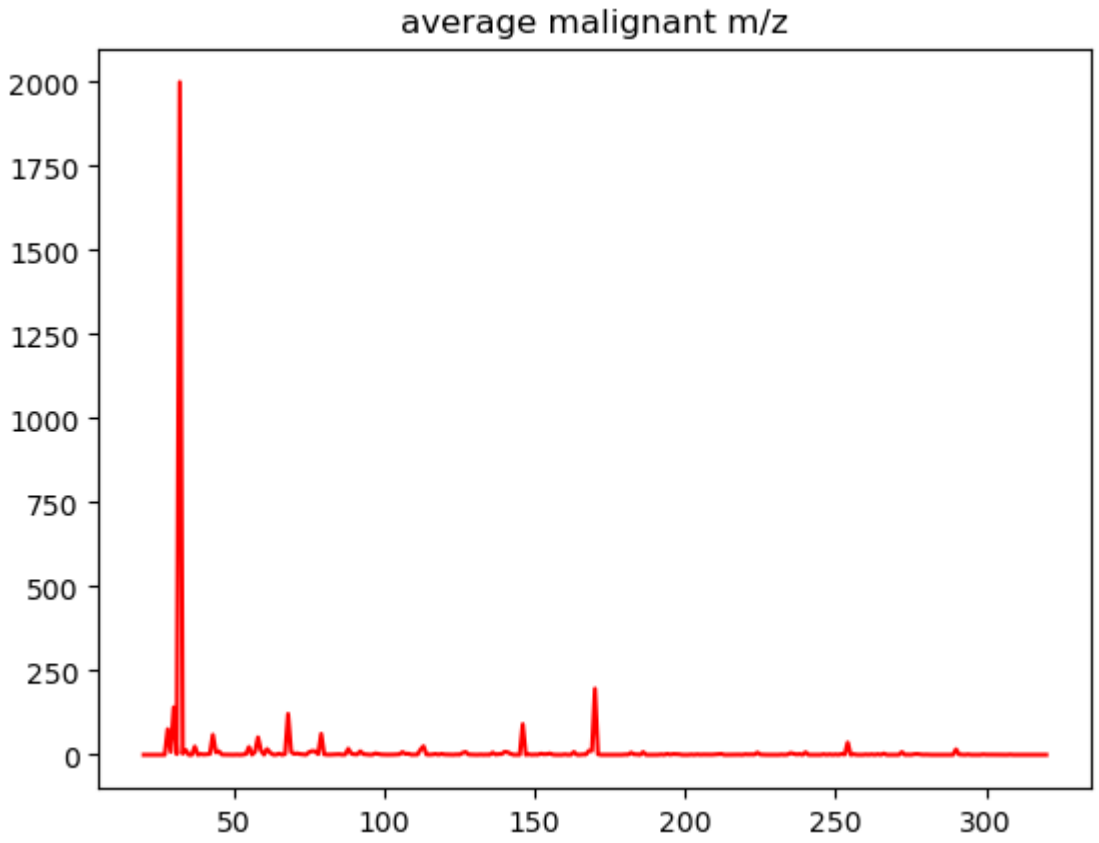

Comorbidity proportion by label
Comorbidity
0
1
proportion
0.0
0.2
0.4
0.6
0.8
Malignant
benign
Label

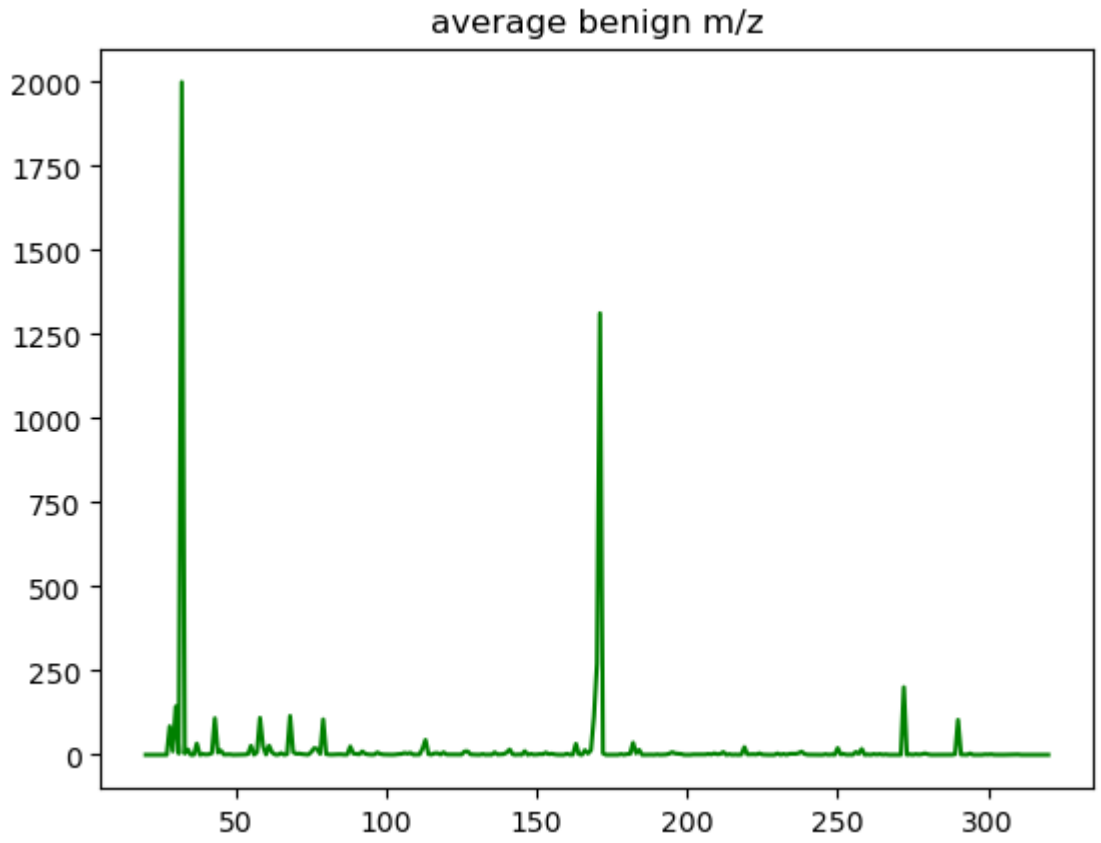

Family history of cancer porportion by label
Family history of cancer
0
1
proportion
0.0
0.2
0.4
0.6
0.8
1.0
Malignant
benign
Label


Drinking history (current/former/never) proportion by label
Drinking history (current/former/never)
0
1
proportion
0.0
0.2
0.4
0.6
0.8
1.0
Malignant
benign
Label

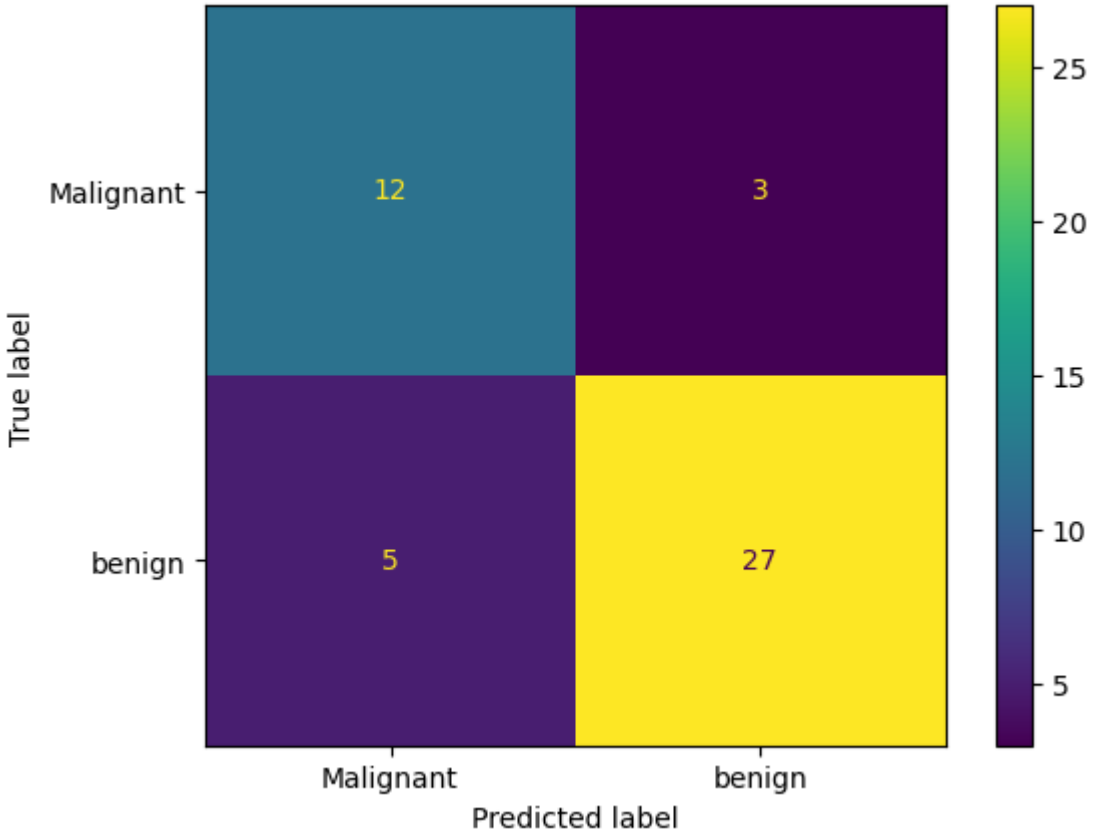

Smoking history (current/former/never) proportion by label
Smoking history (current/former/never)
0
1
proportion
0.0
0.2
0.4
0.6
0.8
1.0
Malignant
benign
Label

EDA


average malignant m/z
2000
1750
1500
1250
1000
750
500
250
0
50
100
150
200
250
300


average benign m/z
2000
1750
1500
1250
1000
750
500
250
0
50
100
150
200
250
300

Results



m/z	Potential VOCs	CAS number	Molecular weight	Molecular formula	feature importance	p-value
105	Pentanethiol	110-66-7	104.214	C5H12S(+H)	0.047455	6.505117e-08
59	Acetone	67-64-1	58.079	C3H6O	0.01863	3.044805e-11
31	formaldehyde	50-00-0	30.031	CH_2O	0.012311	1.523575e-04
116	Isobutyl acetate	110-19-0	116.158	C6H12O2	0.011847	6.691333e-24
42	Acetonitrile	75-05-8	41.05	C_2H_3N	0.011263	7.104396e-17

Conclusion

- **Breath Biopsy® OMNI®:** We identified that the total number of features detected on breath that have more than 100000 and an average of more than mean + 3 * std of blanks is 515 VOCs. Subsequently, we plotted the percentage of features detected on breath samples, where over 350 VOCs appear 50% of the time after filtering. Lastly, we plotted the relative standard deviation for VOCs detected on breath in >50% of samples, where the median inter-subject RSD across these VOCs was 61.84
- **A Clinical Breathomics Dataset:** The Random forest achieved an average accuracy of 95.83% across 5-fold. By utilising SHAP values, we have identified the main 5 VOCs from individuals with asthma, bronchiectasis, and chronic obstructive pulmonary disease. They proved to be better biomarkers than physical descriptors such as sex, age, FVC PP, FEV10 PP, BH (cm) BW (kg), and BMI.
- **Breast cancer-related VOCs:** The Random forest achieved an accuracy of 85.72%, 87.65% Precision, 92.28% Recall and 89.66% F1-score across 5-fold. By utilising SHAP values, we have identified the main 5 VOCs from individuals with malignant breast cancer. These are more relevant than patient characteristics (cancer history, family cancer history, drinking history, comorbidities, smoking history, BMI, etc), however age is still a very strong predictor.
- Models could be improved by exploring more complex models such as ensembled models such as XGBoost. A different feature engineering, such as different scaling or ratios among the variables, could improve results as well. Expanding the datasets sizes with bigger population samples will make these models more robust and reliable to deploy in a live diagnostics server.
- Lastly, the results of this analysis should be compared with larger population sizes, after comparing the VOCs recorded with actual laboratory results of real-life scenarios. Additionally, as we have not recorded the data ourselves, there might be errors unbeknownst to us regarding data quality, thus we advice not to utilise the findings of our analysis as a source of truth.